OUTLOOK
Irritable bowel syndrome

nature

THE INTERNATIONAL JOURNAL OF SCIENCE

Anatomy of the bacteriophage T4 genome-injecting machine PAGE 346



A TRICK OF THE TAIL

NEW MATERIALS

POTENT MIXTURES

Supermix alloys are stronger, tougher, stretchier

PAGE 306

BIOMEDICINE

HOW TO USE STEM CELLS

ISSCR 2016 guidelines for an ethical research framework

PAGE 311

NEW ANTIBIOTICS

UNNATURAL ADVANTAGE

Fully synthetic macrolides to counter resistant microbes

PAGES 326 & 338

NATURE.COM/NATURE
19 May 2016 £10
Vol. 533, No. 7603

THIS WEEK

EDITORIALS

DATABASES Clinical-trial efforts should enforce registration rules **p.292**

WORLD VIEW Why we pushed Mendel off the curriculum **p.293**



GENETICS Shifting sands maintain ancient camel genome **p.295**

Zika must remain a high priority

Although the evidence suggests that the Olympic Games are safe to proceed, the global health community must not let the Zika virus fade from the research agenda until the threat is wiped out.

Are this summer's Olympic Games under threat from the Zika virus? Or, more importantly perhaps, are the competitors and spectators? Athletes such as US women's soccer goalkeeper Hope Solo have said they will go "begrudgingly", and will barely leave their hotel rooms. Some have called for the event to be cancelled.

This overestimates the risk of Zika to visitors. Although public-health agencies have advised pregnant women to avoid countries with active Zika transmission owing to the threat of birth defects, there are much more pertinent threats to the average visitor to Brazil, including dengue virus and random street violence. Much remains unknown about Zika, but a great deal is known, too, and it suggests that there is no reason to cancel the event.

That we know so much so soon is a victory. When the US Centers for Disease Control and Prevention (CDC) concluded last month that the Zika virus causes birth defects, that marked the conclusion of one of the quickest-ever basic-science investigations into a crucial public-health issue. Some researchers expected it to take years. In the end, it took just six months.

The CDC cited a growing number of studies using a range of approaches, including epidemiological, molecular and pathological, that showed a link between Zika and birth defects. The wisdom of making such a declaration so quickly remains a matter of some debate, but it is noteworthy that the global scientific community was able to organize itself in a relatively rapid fashion. There are three main reasons why this happened, and they offer lessons for research in future outbreaks.

First, Brazil has a substantial amount of research infrastructure. Unlike some regions where outbreaks occur, such as nations in Africa, Brazil has made sizeable investment in research and public health. It is the only country in Latin America that invests more than 1% of its gross domestic product into research, and the country is well supported by foreign research institutes, such as the US National Institutes of Health (NIH). The country spent US$31.9 billion on research in 2013, for instance.

This meant that when the Zika outbreak arose, there was a highly trained group of researchers ready to investigate the issue. The Oswaldo Cruz Foundation, a federally funded network of research stations across Brazil, has led investigations of the epidemiology and molecular aspects of Zika. And a long-running NIH-funded project on urban slum health in northeastern Brazil quickly adapted to set up large cohort studies to look at the possible link between Zika and birth defects.

Second, there have been significant advances in neuroscience. Some of the most telling evidence for Zika's effects on human brain cells has come from studies of human neural stem cells and organoids — brain-like structures that can be grown in cell culture in the lab. Brain organoids were developed only a few years ago. By using organoids, human stem cells and human fetal tissue, researchers were quickly able to show that Zika preferentially targets and kills human neural precursors in the brain, offering a plausible explanation for a mechanism through which the virus causes birth defects.

Third, the global health community has acknowledged that its dismally slow response to the Ebola epidemic let it grow out of control. Health officials were therefore determined to move more quickly in response to Zika. The World Health Organization declared the outbreak an emergency in February, and in the same month US President Barack Obama requested $1.9 billion for Zika-related activities. Although the US Congress has yet to approve that spending, the CDC said on 13 May that it had found $85 million to distribute to states to help them prepare for the virus.

"There should be no complacency."

There should be no complacency. The quick connection of dots between Zika and birth defects is only the beginning of the necessary mobilization. And we still don't know why only some babies succumb to the devastating effects of Zika, or the most effective way to care for them. The questions will keep coming, and the research community and funders must continue to pursue them — with speed, strength and the highest urgency, as the organizers of the Olympics might say. ■

Second thoughts

Revisting the past can help to inform ideas of the present.

The thought experiment has a noble place in research, but some thoughts are deemed more noble than others. Darwin and Einstein could let their minds wander and imagine the consequences of certain actions or natural laws. But scientists and historians who try to estimate what might have happened if, say, Darwin had fallen off the *Beagle* and drowned, are often accused of playing parlour games.

Most of these counterfactual thought experiments tend to focus on changes to the lives of historical figures — what would have happened had Hitler never been born, for instance. Dismissed as silly and speculative, such exercises are considered of little academic value, because the results of the experiment tend to align with what the experimenter would have wanted to happen. (One of first such published accounts seems to prove the point: *Napoléon Apocryphe*, published by a supporter in 1841, starts with the emperor surviving the 1812 Russian winter in an unburnt Moscow, conquering Europe, then Asia, Africa and the Americas, while discovering a new planet and inventing a flying car on the way.)

© 2016 Macmillan Publishers Limited. All rights reserved.

The course of history is surely contingent on the roles of influential individuals, which is why counterfactual tales of thwarted plots succeeding and dictators killed as infants seem so poorly anchored to reality. An individual really can steer events, so a world without that individual is unknowable. But what about the course of science, and the ideas that push it along? Is scientific and technical progress equally contingent on circumstance and personality? Or are discoveries inevitable, and independent of the people who happen to be around to make them?

What if Darwin had toppled overboard before he joined the evolutionary dots? That discussion seems useful, because it raises interesting questions about the state of knowledge, then and now, and how it is communicated and portrayed. In his 2013 book *Darwin Deleted* — in which the young Charles is, indeed, lost in a storm — the historian Peter Bowler argued that the theory of evolution would have emerged just so, but with the pieces perhaps placed in a different order, and therefore less antagonistic to religious society.

On page 293, another historian offers an alternative pathway for science: what if the ideas of Gregor Mendel on the inheritance of traits had been challenged more robustly and more successfully by a rival interpretation by the scientist W. F. R. Weldon? Gregory Radick argues that a twentieth-century genetics driven more by Weldon's emphasis on environmental context would have weakened the dominance of the current misleading impression that nature always trumps nurture.

Unusually for a historian, Radick has some experimental data to draw from. Over a term, students at the University of Leeds, UK, where Radick teaches, were presented with a curriculum that relegated Mendel to the margins and promoted Weldon in his place. The result, perhaps not surprisingly, was a class of undergraduates who were less willing to see genetics as immutable destiny.

"Science without consensus would be chaos."

Biologists may take issue with the methods, but the results seem less important than the fact that such an experiment could be performed at all. If the past is a foreign country, then it is also supposed to be one that cannot be revisited. With a little imagination, what other thought experiments could be tested in this way? The history of science is, after all, littered with major theories that became scientific orthodoxy but initially attracted a great deal of tough-minded criticism, from the Newtonian theory of universal gravitation to the Copenhagen interpretation of quantum mechanics.

These 'winners' became dominant before all the criticisms against them were fully answered, which raises questions about why the debates went the way they did, and whether they could have gone otherwise — and if so, with what repercussions.

A well-informed interest in alternative scientific pasts can help us to take the actual past more seriously as a source of present-day insight. It can also help us to stay self-critical as we make choices in the present. Science without consensus would be chaos. But the price of consensus is eternal vigilance against complacency, and a willingness to contemplate the road otherwise not travelled. ■

Open medicine

Governments need to tighten regulation if the sharing of clinical-trial data is to succeed.

Clinical science has a compatibility problem. Although there are set protocols to test medicines and to treat patients, no such standards exist to compare clinical-trial data.

The problem arises because each research group has a preferred method of collecting and categorizing results. Differences can be as great as omitting or including the gender and ethnicity of patients enrolled, or as mundane as the vocabulary used in medical records. For example, a study published in *The BMJ* this week challenged the 'weekend effect' — the idea that people in the United Kingdom admitted to National Health Service hospitals at weekends are more likely to die compared with those admitted on weekdays — by saying that stroke patients' conditions are frequently miscoded by UK hospital staff (L. Li *et al. Br. Med. J.* **353,** i2648; 2016). If proved, such a finding could cast doubt on research studies that use medical records as a data source.

The lack of a single place to report descriptive metadata compounds the problem. Data sharing is most challenging for developing countries, which often lack the resources for large-scale statistical analysis (L. Merson *et al. N. Engl. J. Med.* http://doi.org/bhmb; 2016).

The latest attempt to address this problem is Vivli, described last week in *The New England Journal of Medicine* as a universal platform to "link existing data-sharing platforms and communities, while hosting data from investigators who aspire to share data but lack the resources to do so" (B. E. Bierer *et al. N. Engl. J. Med.* http://doi.org/bhmc; 2016).

The platform, to be run by the Multi-Regional Clinical Trials Center of Brigham and Women's Hospital and Harvard in Cambridge, Massachusetts, will initially employ curators to convert the clinical-trial data they receive from researchers into a standard, anonymized format and post it on the platform for other researchers to access on request. Eventually, the system's creators hope to switch to automated curation.

Vivli's creators acknowledge its limitations. It will no doubt be useful for researchers to verify findings, avoid duplicating trials and prioritize next steps. Similar efforts by scientific societies and researchers on specific diseases have made progress in these directions. But data anonymization and standardization is time-consuming, and so only limited amounts of medical data will at first be available.

Yet the best database in the world cannot address the underlying and pernicious deficiency in clinical-trial data sharing. Unless government regulation is rapidly and substantially changed, Vivli will lack data from the majority of clinical trials — leaving it far from the comprehensive database that its creators envisaged. Despite US laws requiring trial data to be publicly disclosed in the ClinicalTrials.gov database, lax enforcement, loopholes and ambiguous interpretation have allowed pharmaceutical companies to selectively report positive results, alter endpoints and avoid publishing data on time.

Most trials disappear altogether. Some studies estimate that only one-third of trials for drugs approved by the US Food and Drug Administration (FDA) are ever published (J. S. Ross *et al. PLoS Med.* **6,** e1000144; 2009). One assessment found that only two of the ten large pharmaceutical companies complied with FDA regulations on data reporting for drugs approved in 2012 (J. E. Miller *et al. BMJ Open* **5,** e009758; 2015). Gilead and Sanofi Pharmaceuticals were the worst offenders according to author Jennifer Miller of Bioethics International. And private industry is not the only offender — about one-third of federally funded trials remain unpublished four years after completion (J. S. Ross *et al. Br. Med. J.* **344,** d7292; 2012).

Hidden clinical-trial data have long been an open secret in the field, but researchers are now starting to assess it more quantitatively. On 17 May, the Laura and John Arnold Foundation awarded Bioethics International a US$3.6-million grant, which will allow the organization to assess trial-data publication, or lack of it, for all new FDA-approved drugs and biologics.

Ideally, knowledge of such poor compliance will shock the public, pushing the FDA and other regulators into creating stronger incentives for companies to publish complete data, such as serious fines and penalties. Without such intervention, well-intentioned efforts such as Vivli are doomed to remain tragically incomplete, robbing researchers of the opportunity to truly link their work. ■

➲ NATURE.COM
To comment online, click on Editorials at:
go.nature.com/xhunqv

© 2016 Macmillan Publishers Limited. All rights reserved.



SIMON AND SIMON PHOTOGRAPHY

Teach students the biology of their time

An experiment in genetics education reveals how Mendel's legacy holds back the teaching of science, says **Gregory Radick**.

Historians study the causes and consequences of past events, but also consider alternative scenarios. What might have happened, for example, if Britain had stayed out of the war in Europe in 1914? Science historians also ask such counterfactual questions, and the results can be surprisingly instructive.

Take genetics. The past year has seen prolonged celebrations of the work of Gregor Mendel, linked to the 150th anniversary of the paper that reported his experiments with hybrid peas. Mendel's experiments are central to biology curricula across the world. By contrast, the criticisms levelled at Mendel's ideas by W. F. R. Weldon, Linacre professor at the University of Oxford, UK, are a footnote.

From 1902, Weldon's views brought him into increasingly bad-tempered conflict with Mendel's followers. In basic terms, the Mendelians believed that inherited factors (later called 'genes') determine the visible characters of an organism, whereas Weldon saw context — developmental and environmental — as being just as important, with its influence making characters variable in ways that Mendelians ignored. The Mendelians won — helped by Weldon's sudden death in 1906, before he published his ideas fully — and the teaching of genetics has emphasized the primacy of the gene ever since.

The problem is that the Mendelian 'genes for' approach is increasingly seen as out of step with twenty-first-century biology. If we are to realize the potential of the genomic age, critics say, we must find new concepts and language better matched to variable biological reality. This is important in education, where the reliance on simple examples may even promote an outmoded determinism about the power of genes.

But what if Mendelism had never come to dominate genetics in the first place? What if Weldon's perspective had emerged as the winner in that historical battle, and his interactionism, allied to his vivid sense of how variable the real characters of real organisms are (never just yellow or green, round or wrinkled, or any other Mendelian binary), had become the core of the subject? This is where I, and colleagues, have tried to run an experiment.

In a recent two-year project, we taught university students a curriculum that was altered to reflect what genetics textbooks might be like now if biology circa 1906 had taken the Weldonian rather than the Mendelian route. These students encountered genetics as fundamentally tied to development and environment. Genes were not presented to them as what inheritance is 'really about', with everything else relegated to ignorable supporting roles. For example, they were taught that although genes can affect the heart directly, they also affect blood pressure, the body's activity levels and other influential factors, themselves often influenced by non-genetic factors (such as smoking). Where in this tangle, we ask them, is a gene for heart disease? In effect, this revised curriculum seeks to take what is peripheral in the existing teaching of genetics and make it central, and to make what is central peripheral.

Our experimental group consisted of second-year humanities undergraduates. First-year biologists, who were taught the conventional approach, acted as our control. We saw a difference — those students taught the Weldon way emerged as less believing of genetic determinism, and, I suspect, better prepared to understand the subtleties of modern genetics. (The difference was statistically significant, but I hesitate to make much of that, given that numbers were small and there were differences between the groups. I am mindful, too, that it was Weldon who first drew attention to Mendel's own problems with exaggerated statistics.)

With such experiments — bringing insights from the archive into the science classroom — the scientific past can inform and maybe even improve the scientific future. In turn, they suggest a broader vision of collaboration. To advance scientific knowledge, historians and philosophers of science should work in close proximity to scientists, not actually in the lab but right down the corridor. Then, investigations into neglected phenomena and debates that were shut down too soon might provide the spark to serve creative science.

What of Mendel? Some might complain that it is a poor anniversary gift to jettison him from his place of honour in the genetics curriculum. Let me suggest that this grumbling, although understandable, is misguided. If we want to honour Mendel, then let us read him seriously, which is to say historically, without back-projecting the doctrinaire Mendelism that came later. Study Mendel, but let him be part of his time.

Likewise, let our biology students be part of their time, by giving them a genetics curriculum fit for the twenty-first century. If we teach them about Mendel, we should do so not to fill them with slack-jawed wonder at his foundational achievement, but to help them to appreciate how even the most imaginative and rigorous science — and Mendel's was first rate on both counts — bears the stamp of the historical circumstances of its making. To learn that lesson about past science is to bring a welcome level of self-awareness and critical self-reflection to the present. ■

EVEN THE MOST IMAGINATIVE AND RIGOROUS SCIENCE BEARS THE STAMP OF THE HISTORICAL CIRCUMSTANCES OF ITS MAKING.

➲ NATURE.COM
Discuss this article online at:
go.nature.com/afzaos

Gregory Radick *is professor of history and philosophy of science at the University of Leeds, UK, and president of the British Society for the History of Science. This is an edited version of his presidential address, to be published in the* British Journal for the History of Science.
e-mail: g.m.radick@leeds.ac.uk

© 2016 Macmillan Publishers Limited. All rights reserved.

RESEARCH HIGHLIGHTS

Selections from the scientific literature

DIAGNOSTICS

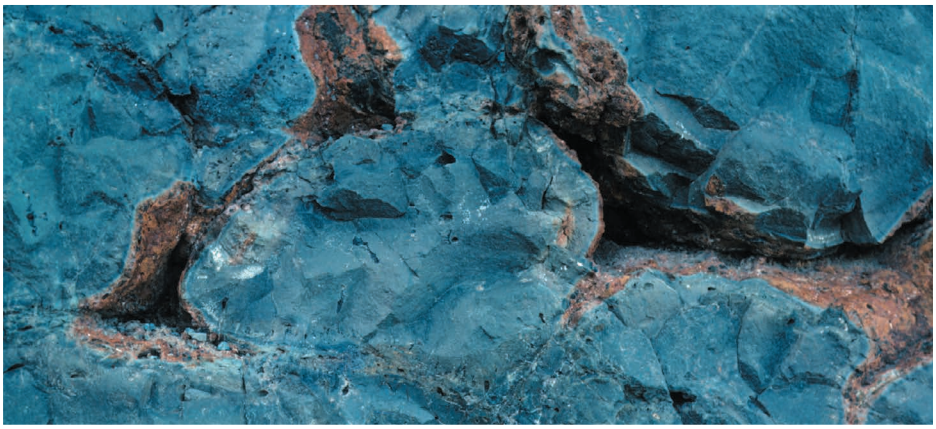
Fast Zika test developed

A low-cost paper-based test can rapidly detect low levels of Zika virus in human blood samples.

James Collins at the Massachusetts Institute of Technology in Cambridge and his colleagues previously designed a paper-based test for Ebola virus. Specially programmed RNA 'sensors' embedded in paper strips bind to viral RNA sequences, triggering a chemical reaction that changes the colour of the paper. The researchers adapted this for the Zika virus, and added a step that boosts the amount of viral RNA, increasing the test's sensitivity. They detected Zika virus in human serum spiked to mimic samples from infected patients. The test also distinguished Zika from the related dengue virus, and can differentiate between two similar strains of Zika.

It takes about 5 days to design the RNA sensors and produce the paper strips, and roughly 3 hours to run the test, making the system potentially useful in future disease outbreaks, the authors say.

Cell **http://doi.org/bhjf (2016)**



DON FRANCIS

GEOCHEMISTRY

Ancient Earth rocks emerge from lava

Deep inside Earth are rocks that date back to the first 50 million years of the Solar System's history. The discovery suggests that, even as internal heating mixes most of Earth's interior, some primordial material still survives today.

A team led by Hanika Rizo of the University of Quebec in Montreal, Canada, analysed lava that erupted off the coast of northern Canada and the southwest Pacific Ocean, and then cooled to form basalt rock (pictured). Both places are thought to tap deep and ancient sources of lava.

The scientists found that levels of tungsten-182 — an isotope created early in Earth's history by the radioactive decay of hafnium-182 — were higher in the Canadian and Pacific rocks than in those from other parts of the world, confirming that the rocks contained ancient materials.

Science 352, **809–812 (2016)**

ASTROPHYSICS

Model predicts neutron-star signal

Physicists have devised a fast, accurate model that recreates the gravitational-wave signals produced by spiralling and colliding neutron stars. The model could help researchers to work out the stars' properties.

The US Advanced Laser Interferometer Gravitational-Wave Observatory (LIGO) is looking for gravitational waves — ripples in space-time generated by cataclysmic cosmic events, such as collisions of stars and other bodies. To learn more about neutron stars from these signals, researchers need a wide range of simulated examples of collisions for comparison. Creating a full simulation can take several weeks, so Tanja Hinderer at the University of Maryland in College Park and her colleagues produced a simplified model. By showing in detail how the stars deform under each other's strong gravitational pull, the model predicts waveforms that agree better with full simulations than do previous such models.

The authors plan to use their model to create thousands of different possible waveforms that will help researchers to interpret LIGO data.

Phys. Rev. Lett. 116, **181101 (2016)**

GENETICS

Genes linked to different traits

Seemingly unrelated biological traits, such as the risk of schizophrenia and of inflammatory bowel disease, can be linked to the same gene variants.

One version of a gene can affect multiple traits, but identifying whether a specific gene variant affects two unrelated pathways has been difficult. Joseph Pickrell of the New York Genome Center and his team analysed data from 43 genome-wide association studies (GWAS) that correlated gene variants with traits such as height, nose size and neurological disorders. They found several clusters of traits that shared common genetic roots, such as one associated with lower body-mass index, increased height and reduced risk of male pattern baldness.

Considering data from many traits is important when studying the effect of a genetic variant, the authors say.

Nature Genet. **http://dx.doi.org/10.1038/ng.3570 (2016)**

ASTRONOMY

Black hole weighed with precision

Astronomers have made precise measurements of the mass of a supermassive black hole.

Aaron Barth of the University of California in Irvine focused the Atacama Large Millimeter/

© 2016 Macmillan Publishers Limited. All rights reserved.

submillimeter Array (ALMA) in Chile on the black hole at the heart of NGC 1332, a galaxy that is roughly 22 million parsecs (73 million light years) away. ALMA's high-resolution images revealed carbon monoxide gas orbiting just outside the black hole. By measuring the velocity of the gas, the team found the black hole to be 664 million times the mass of the Sun, with an uncertainty of roughly 10%.

The technique improves on less-direct measurement methods.

Astrophys. J. Lett. **in the press; preprint at http://arxiv.org/abs/1605.01346 (2016)**

EVOLUTION

Commerce kept camel DNA diverse

RONALD WITTEK/GETTY

Ancient caravans travelling through the deserts of the Middle East and Africa over the past several thousand years have shaped the genetic diversity of domestic camels.

To learn about camels' evolutionary history, a team led by Pamela Burger at the University of Veterinary Medicine in Vienna and Olivier Hanotte at the University of Nottingham, UK, analysed DNA from the remains of 15 ancient dromedaries and from more than 1,000 living ones from 21 countries. They found that domestic camels (*Camelus dromedarius*; **pictured**) are most closely related to wild ones now living in the southeast Arabian Peninsula, and that their DNA varies little by geography — a possible relic of ancient trading routes that mixed the genes of distant populations.

Unlike some other domesticated livestock, camels seem to have largely maintained their ancestral genetic diversity.

Proc. Natl Acad. Sci. USA **http://doi.org/bhf7 (2016)**

DMVALEY/GETTY



STEM CELLS

Self-help for type 1 diabetes

Cells from people with type 1 diabetes can be reprogrammed so that they produce insulin in response to glucose when implanted into mice. This suggests that patients with the disease could one day be treated with their own reprogrammed cells.

In previous work, Douglas Melton at Harvard University in Cambridge, Massachusetts, and his team took connective-tissue cells from healthy donors, reprogrammed them into stem cells and then grew β-cells, which make insulin in the pancreas. Now the team has done this using connective-tissue cells from people with type 1 diabetes. The β-cells derived from these patients were similar to those derived from healthy individuals — they prevented diabetes in mice that had lost their own β-cells, and responded to anti-diabetes drugs.

Nature Commun. 7, **11463 (2016)**

CELL BIOLOGY

'Nanoblade' slips organelle into cells

A laser-based technique permits the delivery of energy-generating organelles called mitochondria into single mammalian cells, where they can restore metabolic activities.

Mitochondria have their own DNA, which can cause disease if mutated, and researchers have been looking for ways to isolate the organelles, correct genetic defects and return the mitochondria to cells. Pei-Yu Chiou and Michael Teitell at the University of California, Los Angeles, and their colleagues coated a micropipette tip with a 100-nanometre-thick film of light-absorbing titanium and lightly touched it to a cell membrane. A laser pulse heated the tip, causing a bubble to quickly form and collapse in the cell's culture medium. The bubble's expansion punctured the membrane, creating an opening large enough for the delivery of mitochondria.

The technique's 2% efficiency rate is higher than that of other methods, and the team is now working to increase throughput.

Cell Metab. 23, **921–929 (2016)**

SOCIAL SELECTION

Popular topics on social media

Epigenetics article kicks up a storm

A story in the 2 May issue of *The New Yorker* has been criticized for inaccurately describing how genes are regulated. The article by Siddhartha Mukherjee — a physician, cancer researcher and award-winning author at Columbia University in New York — examines how environmental factors change gene activity without altering DNA sequence. Jerry Coyne, an evolutionary ecologist at the University of Chicago in Illinois, wrote two blog posts calling the piece "superficial and misleading", largely because it ignored key aspects of gene regulation. Other researchers quoted in the blog called it "horribly damaging" and "a truly painful read". Mukherjee responded with a point-by-point rebuttal online. Speaking to *Nature*, he says he now realizes that he erred by omitting key areas of science, but that he didn't mean to mislead. "I thought that I had done it justice," he says.

➲ **NATURE.COM**
For more on popular papers:
go.nature.com/rnob6u



ECOLOGY

Wolf cull makes poaching worse

Government-approved killing of wolves (**pictured**) increases illegal hunting in parts of the United States.

State-sanctioned culls are thought to be an effective conservation tool for reducing poaching of large carnivores. Guillaume Chapron at the Swedish University of Agricultural Sciences in Riddarhyttan and Adrian Treves at the University of Wisconsin–Madison studied wolf populations in Michigan and Wisconsin between 1995 and 2012. During that time, policy shifts meant that wolves experienced periodic stages of protection and legal culling. The authors found that population growth slowed during periods of culling, regardless of the number of wolves culled.

The authors ruled out other explanations for the decline, and suggest that culling increases people's intolerance of endangered predators, leading to more poaching. They recommend a moratorium on culling until its effects are better understood.

Proc. R. Soc. B 283, **20152939 (2016)**

➲ **NATURE.COM**
For the latest research published by *Nature* visit:
www.nature.com/latestresearch

© 2016 Macmillan Publishers Limited. All rights reserved.

SEVEN DAYS *The news in brief*

POLICY

Methane rules

The US Environmental Protection Agency finalized a trio of regulations on 12 May that aim to reduce emissions of methane and other pollutants from oil and gas fields. Methane is 25 times more powerful than carbon dioxide at trapping heat in the atmosphere, and the agency projects that the new regulations will reduce methane emissions by the equivalent of 11 million tonnes of CO_2 in 2025. The regulations apply only to new and modified operations, but the agency says that it is planning a separate regulatory process for existing ones.

Costly climate

The annual cost of adapting to a warming world for developing countries could reach US$280 billion to $500 billion by 2050, according to the United Nations Environment Programme's second Adaptation Gap Report, released on 10 May. Countries and multilateral institutions have increased adaptation aid, which totalled $22.5 billion in 2014, according to the report, but adaptation costs are already two or three times higher than that. The agency also called for a system to track investments and assess progress in meeting goals.

RESEARCH

New worlds galore

NASA's Kepler space telescope has in one stroke more than doubled the list of planets it has discovered around other stars, thanks to a new method of validating signals of possible worlds. On 10 May, scientists confirmed the existence of 1,284 new extrasolar planets from a pool of 4,302 candidates that Kepler has spotted since 2009. They include about 100 roughly Earth-sized bodies and nine planets in the habitable zones of their stars. The announcement raises the total number of confirmed extrasolar planets to 3,264. See go.nature.com/mxwalg for more.



FLIP NICKLIN/ MINDEN PICTURES/NATIONAL GEOGRAPHIC CREATIVE

Porpoise on brink of extinction

The vaquita (*Phocoena sinus*), a porpoise unique to the Gulf of California off Mexico, could be extinct by 2022, according to a report released on 14 May by the International Committee for the Recovery of the Vaquita (CIRVA). On the basis of acoustic data from submarine recorders and a survey of the vaquita's entire range, CIRVA estimates that there are now fewer than 60 animals left. Its dramatic decline of 92% in just under two decades is attributed to the use of gill nets in fishing. Mexico implemented a two-year gill-net ban in 2015, but some fishermen have continued to use the nets, resulting in the deaths of at least three vaquitas this year. At a meeting in Ensenada, Mexico, last week, CIRVA scientists called on the Mexican government to make the gill-net ban permanent.

Poor air quality

The vast majority of city dwellers in poorer countries endure unsafe levels of air pollution, according to an updated database released by the World Health Organization (WHO) on 12 May. In low- and middle-income countries, 98% of people in large cities that monitor outdoor pollution breathe air that exceeds the limits set by the WHO, according to the database, which now covers 3,000 towns and cities in 103 countries. Among megacities with more than 14 million people, Delhi, Cairo and Dhaka experienced the worst levels of pollution from fine particles, which increase the risk of stroke, heart disease, lung cancer and asthma. Globally, urban air pollution worsened by 8% between 2008 and 2013, despite improvements in Europe and the Americas.

FUNDING

Cosmic funding

The Simons Foundation in New York will contribute the majority of a US$40-million donation to establish an observatory to look for primordial gravitational waves, the non-profit organization announced on 12 May. The Simons Observatory will scour the Universe's all-pervading background radiation for imprints of the waves. Finding them would provide evidence of a theorized period of rapid inflation in the moments after the Big Bang. (The BICEP2 experiment made a claim to

© 2016 Macmillan Publishers Limited. All rights reserved.

have discovered this imprint in 2014, but subsequently retracted it.) The telescopes will be built in Chile's Atacama Desert and aim to be around ten times more sensitive than any currently operating facility.

Funding rejig

Britain's seven research councils are to be brought together into one 'supercouncil', according to a government White Paper unveiled on 16 May. The overarching body will be called UK Research and Innovation, and will also include the agency Innovate UK and the research-funding portions of the Higher Education Funding Council for England. Under the plan, the agencies will retain their own identities and separate budgets. The plan also calls for the creation of an Office for Students to oversee higher education, and introduces the Teaching Excellence Framework, which will evaluate teaching at UK universities and allocate funding on the basis of quality.

PEOPLE

Ilkka Hanski dies

Researchers around the world have been paying tribute to Finnish ecologist Ilkka Hanski (**pictured**), who died on 10 May, aged 63. Hanski won the Crafoord



MARKUS MARCETIC/ROYAL SWEDISH ACADEMY OF SCIENCES

Prize in Biosciences from the Royal Swedish Academy of Sciences in 2011 for his work on how spatial variation affects animal and plant populations. In 1991, he founded the Metapopulation Research Centre at the University of Helsinki to study species in fragmented landscapes.

Academy head quits

Biochemist Wong Chi-huey resigned on 10 May as the head of Academia Sinica, Taiwan's foremost research institution. After initially rejecting Wong's resignation in March, Taiwan's President Ma Ying-jeou accepted it following a controversy over Wong's relationship with Taiwanese biotechnology company OBI Pharma. Wong's daughter reportedly sold shares in OBI before their value dropped following the release in February of disappointing results of a trial for a breast-cancer drug. Wong denied any allegations of insider trading, saying that he had no advance knowledge of the trial results.

BUSINESS

Pharma deal

Pharmaceutical giant Pfizer announced a US$5.2-billion deal on 16 May to acquire Anacor Pharmaceuticals of Palo Alto, California, which specializes in boron-based compounds. Pfizer cited crisaborole, an experimental drug for the treatment of eczema that is under review for approval by the US Food and Drug Administration, as Anacor's "flagship asset". In April, Pfizer scrapped plans for a $160-billion merger with Dublin-based Allergan after the US treasury tightened up rules on companies moving their headquarters abroad to cut their tax bills.

Execution drugs

Drug multinational Pfizer will no longer sell drugs for use in executions by lethal injection, the company announced on 13 May. In banning the distribution of 7 drugs to correctional facilities, Pfizer joins 24 other companies that have ended such sales. "Pfizer makes its products to enhance and save the lives of the patients we serve," the company said in a statement. All the makers of such drugs that are approved by the US Food and Drug Administration have now halted sales for executions, according to the London-based group Reprieve, which campaigns against the death penalty.

Preprint purchase

The Social Science Research Network (SSRN), one of the world's most popular repositories of preprint research in economics, law and the social sciences, has been bought by academic-publishing giant Elsevier for an undisclosed sum. After the acquisition, which was announced on 17 May, the SSRN will continue to offer free submissions and downloads, said its chief executive, Gregg Gordon. Social Science Electronic Publishing, a private firm in Rochester, New York, founded the SSRN in 1994. See go.nature.com/5jnxqv for more.

COMING UP

23–24 MAY
The biotechnology conference Big Data Leaders Europe takes place in London.
go.nature.com/uchnje

24–26 MAY
The Thirty Meter Telescope Science Forum, the facility's annual planning meeting, gathers in Kyoto, Japan.
go.nature.com/sbnhua

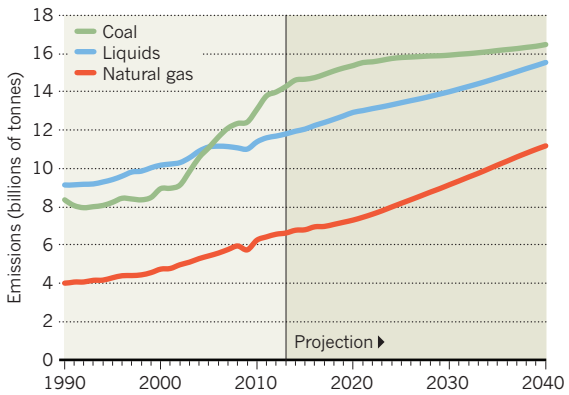
24 MAY
The 13th and 14th satellites of Europe's Galileo navigation system are due to launch from the Guiana Space Centre.
go.nature.com/kqdwgk

TREND WATCH

If current policies and regulations remain, energy-related carbon dioxide emissions may rise by 34% to 43 billion tonnes in 2040. In its *International Energy Outlook 2016*, released on 11 May, the US Energy Information Administration (EIA) projects that energy consumption will rise by 48% by 2040, spurred by growing demand in Asia. Although renewable energy is the fastest-growing energy sector, fossil fuels are likely to continue to supply more than three-quarters of the world's energy, the EIA said.

SOBERING OUTLOOK

Carbon dioxide emissions from fossil fuels will continue to rise for decades unless climate goals are pursued with a vengeance.



SOURCE: EIA

NATURE.COM
For daily news updates see:
www.nature.com/news

© 2016 Macmillan Publishers Limited. All rights reserved.

NEWS IN FOCUS

POLITICS Brazil's new interim president downgrades science ministry **p.301**

BOATY MCBOATFLEET British ship joins global wave of ice-hardy vessels **p.302**

NANOMEDICINE Bankruptcy at trailblazer firm worries drug developers **p.304**



AGRICULTURE The race is on to create fertilizer-efficient crops **p.308**

ALESSIO PADUANO/POLARIS/EYEVINE



A farmer works in southern Italy, where the bacterium *Xylella fastidiosa pauca* is devastating olive trees.

PLANT PATHOGENS

Olive tree gridlock eases

Court ruling paves way for containment of deadly plant disease, but risk of spread remains.

BY ALISON ABBOTT, BARI, ITALY

From a small hill in the Puglia region of southern Italy, plant pathologist Donato Boscia gestures towards a landscape of the dullest brown — dead and dying olive trees as far as the eye can see. Six months earlier, he says, that canopy was mostly green, with just a few tell-tale brown spots marking the relentless advance of a vicious pathogen, *Xylella fastidiosa*, that was previously unknown in Europe.

Almost three years ago, Boscia, who heads the Bari unit of the CNR Institute for Sustainable Plant Protection (IPSP), and other colleagues identified a subspecies of the bacterial pathogen, *Xylella fastidiosa pauca*, as the cause of olive quick decline syndrome (OQDS) in Puglia. They said that it probably arrived with an ornamental plant imported from the Americas, where some *Xylella* species are endemic.

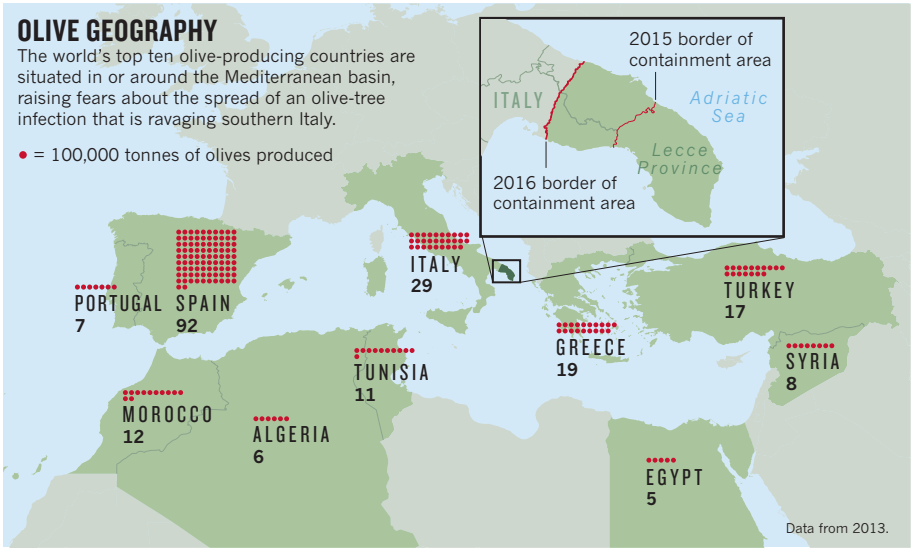
But an alarming political and legal impasse has stopped measures to contain the pathogen, which has invaded nearly 200,000 hectares of olive groves and is killing most olive trees in its wake — including beloved specimens that are more than 1,000 years old.

Now the gridlock shows some signs of easing. On 12 May, the European Court of Justice declared that containment measures agreed by the European Union more than a year ago — but that were quickly blocked by dismayed protesters — were indeed appropriate, paving the way for them to now go ahead. These include monitoring the disease's spread, uprooting infected trees and, in some cases, any apparently healthy trees surrounding them, and killing the insects that carry the bacteria. There is currently no cure for *Xylella*, and evidence from the Americas indicate that it cannot be eradicated.

Still, challenges to fighting the outbreak remain. The infected area has grown over the past year, increasing the chance that the disease will spread to other countries in the Mediterranean basin, which accounts for 95% of the world's olive production (see 'Olive geography').

In February last year, Italian authorities began implementing the EU-agreed containment plan. But protesters, with local political support, dismissed the evidence that OQDS was caused by *Xylella*, and that it cannot be cured or eradicated. They argued that the measures were not warranted and took their case to a local administrative court, which referred it to the European Court of Justice. ▶

© 2016 Macmillan Publishers Limited. All rights reserved.



FAOSTAT/EUROPEAN COMMISSION

▶ Paralysed by uncertainty, local authorities blocked even monitoring efforts.

Aside from their economic importance, olive trees have a long historical and cultural connection with these regions. "So the public resistance is understandable," says plant geneticist Francesco Salamini, who is leading a small group of scientists preparing an independent report on the affair for Italy's national academy.

Confusing perceptions further, last December public prosecutors opened an investigation into whether five scientists — Boscia included — and five public officials negligently spread the disease, deceived officials about it and caused environmental pollution, among other things. IPSP director Gian Paolo Accotto, who is based in Turin, calls these notions "logically absurd".

CONTROL PLAN

The case against containment began to unravel about eight weeks ago when the Parma-based European Food Safety Authority (EFSA), which provides independent scientific advice to the EU, released three reports[1–3] supporting the role of the bacteria in causing the disease and endorsing the EU-agreed containment measures. Puglia regional authorities then approved a new containment plan that defines a greater area of the region as infected (see inset in 'Olive geography'). Early this month, administrators there agreed to restart monitoring. A few days later, the European Court of Justice made its declaration.

The need for containment is highlighted by Boscia's work over the past year inoculating different plant species with *X. f. pauca* in modest greenhouses at the IPSP in Bari. This has revealed that citrus trees and grape vines are immune to the pathogen, but that it can infest a large number of local species such as lavender, oleander and polygala. Researchers say that even if all infected olive trees were destroyed, the bacteria would be so widely harboured in the environment that containment is the best measure.

"You can stop humans from trying to help, but you can't stop bacteria, or put them in jail."

The Bari experiments also revealed that some cultivars of olive tree — including Leccino, the main cultivar in the Tuscany region — developed milder symptoms of disease than do the predominant cultivars of Puglia. This could inform efforts to deal with the pathogen should it spread to other regions. It also helps replanting schemes, although more seasons of growth are required to see whether the Leccino trees survive long-term, says Boscia.

Other experiments at the IPSP are analysing gene expression in infected trees to determine which molecular characteristics might offer some resilience. This could one day inspire treatments for infected trees.

The Bari scientists, together with a group of worried olive producers, are also carrying out experiments in the infected zone. Enzo Manni, head of an association that represents 600 small olive groves there, says that he was stonewalled a few years ago when he tried to interest local institutions in what seemed to be a threatening new disease. "But I knew something was really wrong," he says. Now he helps the Bari researchers by identifying land for outdoor experiments and monitoring the tests.

TOLERANCE TEST

In one of these experiments, the team of scientists and producers — which has now grown to 40 — has planted different cultivars to test whether any show tolerance over many seasons. Such experiments could guide which trees get planted in the region, or elsewhere. In another study, the scientists are grafting different cultivars onto newly infected trees to see whether some survive, thereby offering hope of recreating canopies on diseased root stock. In some cases, they are funding some of these experiments themselves so as not to lose time.

Accotto applauds the scientists' commitment in spite of the police investigations. "You can stop humans from trying to help, but you can't stop bacteria, or put them in jail," he says.

Fearing the spread of *Xylella*, scientists in other countries are looking to join in. Those from Spain, the world's largest producer of olives, have added a couple of the country's cultivars to the experiments to get an idea of how fast their trees might succumb to the disease.

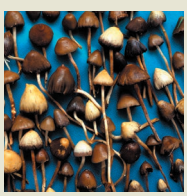
The United Nations Food and Agriculture Organization (FAO), based in Rome, organized an April workshop in Bari to share information with scientists and agriculture officials from olive-growing countries in Europe, North Africa and the Middle East.

The delay to the containment programme in Italy has probably increased the risk of OQDS spreading abroad, says Shoki Al-Dobai, head of the FAO's North Africa and Middle East office in Cairo. He is organizing a contingency plan for that area. Syria and Libya, where olive cultivation is believed to have begun in 2,500 BC, are of particular concern, he says, because wars in those countries make appropriate monitoring impossible. ■

1. Saponari, M. *et al. Pilot Project on* Xylella fastidiosa *to Reduce Risk Assessment Uncertainties* (EFSA, 2016); available at http://go.nature.com/aldezo
2. EFSA Panel on Plant Health. *EFSA J.* **14,** 4450 (2016).
3. EFSA Panel on Plant Health. *EFSA J.* **14,** 4456 (2016).

MORE ONLINE

TOP NEWS



Magic-mushroom drug lifts depression in first human trial go.nature.com/isijvz

MORE NEWS

● LHC for quasiparticles reveals material secrets go.nature.com/84ebtc
● Digital badges motivate scientists to share data go.nature.com/hyjbnm
● Underwater archaeologists unearth ancient butchering site go.nature.com/feazlv

NATURE PODCAST

Treasures from sunken cities; antibiotics from scratch; and experimenting with history nature.com/nature/podcast

DIVERSE IMAGES/UIG/GETTY

© 2016 Macmillan Publishers Limited. All rights reserved.

POLITICS

Brazil's science woes mount

Demotion of science ministry rubs salt in wound following severe federal funding cuts.

BY CLAUDIO ANGELO



ADRIANO MACHADO/REUTERS

Interim President Michel Temer has decided that Brazil will no longer have a separate science ministry.

Brazil's scientists, already struggling to absorb massive funding cuts, are protesting against another blow: the country's science ministry has been demoted by interim president Michel Temer, who took over the government on 12 May after a Senate impeachment vote ousted Dilma Rousseff from the presidency.

One of Temer's first actions was to announce the fusion of the federal Ministry of Science, Technology and Innovation (MCTI) with the ministry that deals with telecommunications and Internet regulation. Science is now an office within a 'superministry' led by Gilberto Kassab, a former mayor of São Paulo.

The move, which Temer had begun to hint at a few days before, has angered some Brazilian researchers. On 11 May, 13 scientific associations, led by the Brazilian Society for the Advancement of Science and the Brazilian Academy of Sciences (ABC), sent a letter to Temer warning that the fusion would be "detrimental to the country's scientific and technological development". Researchers say that it would erode the authority of the MCTI, which has formed the backbone of federal support for science and innovation in Brazil for the past three decades.

"An administrative reorganization should stem from a vision for the country. It shouldn't simply be an artificial lumping of disparate activities," says ABC president Luiz Davidovich. The move is "a step backwards", he says.

Temer had first caused concern among scientists the week before, when he hinted to local media that he might appoint a creationist evangelical bishop to lead the science ministry. "Unfortunately, the science ministry is usually among the first bargaining chips in every new government," says José Eduardo Krieger, dean of research at the University of São Paulo.

Brazil has had three science ministers in the past 16 months — and each was seemingly appointed for political ends rather than for any particular expertise. In January last year, Rousseff picked Aldo Rebelo, an avowed climate-change sceptic in the Communist party (the closest ally of Rousseff's centre–left Worker's Party). He was followed by Celso Pansera, whose nomination was regarded as an attempt to lure votes from Temer's Brazilian Democratic Movement Party against Rousseff's impeachment. And Emília Ribeiro, former vice-minister of the MCTI, has been science minister since April 2016.

Pansera says that such high turnover is detrimental to research. "It gives you a lot of discontinuity," he says.

FUNDING SLASHED

The most hurtful discontinuity for Brazil's researchers has been startling cuts to federal funding. Last year, as the nation's fiscal crisis began to bite, the MCTI's budget was cut by some 1.9 billion reais (US$540 million), to 5.4 billion reais. Its budget for this year was set 37% lower than last year's — and its authorized spending has been chopped by another 6%. The collapse of oil prices and a bribery scandal that involves Petrobras, the behemoth state-run oil company, have further reduced cash flows for Brazilian research, which partly depends on oil revenue.

A major victim of the economic downturn is the flagship exchange programme Science Without Borders, which by the end of 2015 had sent nearly 94,000 Brazilian undergraduate and postgraduate students to institutions abroad. The programme intended to send a further 100,000 students abroad by 2018, but its second phase has been called off.

To revive science spending, the government had authorized the MCTI to negotiate a US$1.4-billion loan from the Inter-American Development Bank in Washington DC. But that operation has been reset, Pansera says. "Now we'll have to renegotiate everything with the new government's economic team."

Research institutes are struggling to survive. At the Brain Institute at the Federal University of Rio Grande do Norte in Natal, researchers split basic maintenance expenses such as Internet bills among themselves and buy equipment with their own money, says institute director Sidarta Ribeiro.

And some scientists are already planning to leave the country. Suzana Herculano-Houzel, a neuroscientist at the Federal University of Rio de Janeiro, launched a crowdfunding campaign last year to buy new parts for a microscope. But she has now decided to move to Vanderbilt University in the United States: this September, she will shut down her lab in Brazil, which had been running for a decade.

Brazilian scientists are familiar with economic crises. In the 1990s, budget constraints were the rule, recalls Paulo Artaxo, a physicist at the University of São Paulo. But before the current crisis, the country's research had been experiencing unprecedented expansion, with funding on the rise and the number of PhD students soaring. In 2011, Brazil rose to 13th in the world in research-paper production, up from 17th in 2001. That trend is reversing, says Rogério Meneghini, a specialist in science metrics and scientific director of SciELO, a collection of mainly Latin American journals. Brazil's research-paper production grew 16% between 2011 and 2014 — but his figures suggest that it dropped by 4% in 2015. "What is cruel about this is that, when you cut off on science, you can never resume from where you stopped", Krieger says. "You lose position." ■

© 2016 Macmillan Publishers Limited. All rights reserved.



PETE BUCKTROUT/BRITISH ANTARCTIC SURVEY

The RRS *James Clark Ross* is one of Britain's two ageing polar ships, which are being replaced.

OCEAN SCIENCE

Fleet of polar ships raises science stakes

RRS Sir David Attenborough — *named Boaty McBoatface by the Internet — is part of a wave of hardy ice-going vessels.*

BY ALEXANDRA WITZE

The United Kingdom's new polar research vessel has become a national obsession. A proposal to christen the ship Boaty McBoatface captured the public imagination — so much so that the decision to name the ship instead after broadcaster and naturalist David Attenborough triggered questions in a parliamentary inquiry.

But the RRS *Sir David Attenborough*, which will probe both the Arctic and Antarctic, is also notable because it will carry more scientists and push deeper into polar ice than any UK research vessel ever has. It joins a wave of ice-strengthened research vessels — from Norway to China to Australia — that promise to expand scientists' ability to explore harsh polar environments (see 'Ships of the future').

China is soliciting bids to build a large research icebreaker to accompany its existing ship *Xuelong*, and the Australian government signed a contract last month for a replacement for its ageing *Aurora Australis*. Germany has started the process of replacing its vessel *Polarstern*, and Sweden is beginning to discuss what to do after its ship *Oden* is retired towards the end of the next decade. (A proposed pan-European icebreaker, costing up to €800 million (US$900 million), is on hold owing to its enormous price tag.)

The 129-metre-long *Attenborough* is designed by the team behind the *Kronprins Haakon*, an icebreaker being built for the Norwegian Polar Institute in Tromsø. When *Kronprins Haakon* sets off on its first science cruise in 2018, it will give Norwegian researchers a serious upgrade in their access to Arctic seas. It will replace a smaller vessel that can carry fewer scientists and operate only in light ice. "It's a whole new world for us," says project manager Øystein Mikelborg of the Norwegian Polar Institute.

OVER ICE

Within weeks, architects will finalize the *Attenborough*'s design and workers in Merseyside, UK, will begin cutting steel for its hull, on track for a 2019 delivery.

Designers are cramming as many research goodies as possible into the £200-million (US$290-million) ship. The *Attenborough* will have a helicopter deck and, unlike Britain's existing polar-research vessels, a hangar. This will allow scientists to fly to otherwise inaccessible lakes or islands in Antarctica. And the *Attenborough* will have a hole in the hull known as a moon pool, allowing researchers to deploy oceanographic and geological equipment more smoothly and safely than by swinging it off the side of the ship.

Laboratory spaces will be kept at different temperatures, allowing storage and experiments of different types. The deck will have extra room for custom-made equipment for different cruises. Modern fibre-optic cables will deliver a live camera feed from as deep as 6,000 metres, from a remotely operated vehicle that will be called Boaty McBoatface, in a concession to the public vote.

Yet this increased research capability comes at a price: fewer days of actual science.

The British Antarctic Survey (BAS) currently operates two polar vessels, the 25-year-old RRS *James Clark Ross* for science, and the 21-year-old RRS *Ernest Shackleton*, which delivers equipment, food and people to its bases. To cut annual operating costs by 20%, the bigger *Attenborough* will replace both ships.

Instead of 180 days of science a year, UK scientists will get 150. The rest of the time (and space on the ship) will go to ferrying cargo. "I don't like to use the word compromise, but there is a trade-off," says Andrew Jeffries, project manager for the new boat and a BAS engineer in Cambridge, UK.

There have been other trade-offs between research and logistics. Ship designers have had to place the vessel's enormous diesel generators on isolating platforms to make the ship run silently, to avoid disturbing seismic and biological acoustics studies. The current design may also have to be tweaked to avoid vents from the helicopter fuel storage opening directly into areas where atmospheric scientists were counting on having clean air for their measurements.

Even with these trade-offs, researchers are eager for the *Attenborough* to launch. Not only

© 2016 Macmillan Publishers Limited. All rights reserved.

SHIPS OF THE FUTURE

As climate change alters the polar seas, nations are upgrading their ability to conduct research in these ice-strewn waters.

NORWAY *Kronprins Haakon*



100 metres

Ice-breaking capability | ~1 metre
Launch date | 2017
Cost | 1.4 billion kroner (US$170m)

Approxiate number of scientists on board



35

CHINA Unnamed



122.5 metres

Ice-breaking capability | 1.5 metres
Launch date | 2019
Cost | Undisclosed



49

UNITED KINGDOM RRS *Sir David Attenborough*



129 metres

Ice-breaking capability | 1 metre
Launch date | 2019
Cost | £200m (US$290m)



60

AUSTRALIA Unnamed



156 metres

Ice-breaking capability | 1.65 metres
Launch date | 2020
Cost | Aus$529m (US$380m)



116

GERMANY Unnamed



Approx. 120 metres

Ice-breaking capability | 1.5 metres
Launch date | 2020
Cost | Unknown



~50

will it be able to push through thicker ice and carry more scientists, but the ship will also be able to explore deeper environments than any other UK research vessel, says Katrin Linse, a deep-sea biologist at the BAS. That ability will allow it to carry out the first direct sampling of the 8.5-kilometre-deep South Sandwich Trench in the southern Atlantic Ocean. "We might find a new ecosystem there that we are not aware of yet," she says. "It's pretty exciting to get a new vessel. That only happens once every 30–40 years." ■

SYNTHETIC BIOLOGY

Enzyme boost for mirror-image life

Polymerase that can copy left-handed DNA marks step forward for looking-glass biochemistry.

BY MARK PEPLOW

It's biochemistry — but not as we know it. Researchers at Tsinghua University in Beijing have created a mirror-image version of a protein that performs two of the most fundamental processes of life: copying DNA and transcribing it into RNA.

The work is a "small step" along the way to making mirror-image life forms, says molecular biologist Jack Szostak of Harvard Medical School in Boston, Massachusetts. "It's a terrific milestone," adds his Harvard colleague George Church, who hopes one day to create an entire mirror-image cell.

Many organic molecules are 'chiral': they can exist in mirror-image forms that cannot be superimposed, like a right-handed and left-handed glove. But life almost always employs one version: cells use left-handed amino acids, and have DNA that twists like a right-handed screw, for instance. In principle, looking-glass versions of these molecules should work together in the same way as normal ones — but they might be resistant to attack by viruses or enzymes that have not evolved in a looking-glass world.

That makes mirror-image biochemistry a potentially lucrative business. One company that hopes so is Noxxon Pharma in Berlin. It uses laborious chemical synthesis to make mirror-image forms of short strands of DNA or RNA called aptamers, which bind to therapeutic targets such as proteins in the body to block their activity. The firm has several mirror-aptamer candidates in human trials for diseases including cancer; the idea is that their efficacy might be improved because they aren't degraded by the body's enzymes. A process to replicate mirror-image DNA could offer a much easier route to making the aptamers, says Sven Klussmann, Noxxon Pharma's chief scientific officer.

THROUGH THE LOOKING-GLASS

Researchers have been making chunks of mirror-DNA for decades, so the Tsinghua team could order much of what it needed for its looking-glass DNA replication attempt from a chemical supplier — a mirror-DNA strand to be copied, mirror-DNA building blocks and a shorter mirror 'primer' strand that could pick up these building blocks in the right order.

The difficult task was to make the mirror-image enzyme that coordinates the copying process, called DNA polymerase. That would need to be synthesized from right-handed amino acids, but commonly used polymerase enzymes have more than 600 amino acids — meaning that they are too big for current synthetic methods.

So the Tsinghua team turned to the smallest known polymerase: African swine fever virus polymerase X, which contains just 174 amino acids. Unfortunately, it is also spectacularly slow — probably because of its small size, says synthetic biologist Ting Zhu, a former graduate student of Szostak's who helped to lead the work. The team made a mirror version of the enzyme and found that, like its natural equivalent, it could extend a mirror-primer consisting of 12 nucleotides (DNA building blocks) to an 18-nucleotide mirror-DNA strand in about 4 hours; and to a 56-nucleotide strand in 36 hours.

When the normal and mirror-image versions of these systems were mixed together in the same test tube, both replication processes worked independently without interference. The mirror-image polymerase could also transcribe mirror-DNA into mirror-RNA — a relatively rare feat for a polymerase — again at a glacial pace. The work is published in *Nature Chemistry* (Z. Wang *et al. Nature Chem.* http://dx.doi.org/10.1038/nchem.2517; 2016).

"For a while mirror-image biochemistry was a non-field. But now it seems very vibrant."

Klussmann says that Noxxon Pharma is interested in pursuing a similar approach with a more efficient enzyme. Indeed, Zhu and his colleagues next hope to build a mirror-image of a more efficient polymerase known as Dpo4, which is built of 352 amino acids.

In their research paper, the Tsinghua researchers also present their work as an effort to investigate why life's chirality is the way it is. This remains mysterious: it may ▶

© 2016 Macmillan Publishers Limited. All rights reserved.

simply be down to chance, or it could have been triggered by a fundamental asymmetry in nature. But Steven Benner, at the Foundation for Applied Molecular Evolution in Alachua, Florida, says it's unlikely that creating a mirror form of biochemical life could shed light on this question. Almost every physical process behaves identically when viewed in a mirror. The only known exceptions, called 'parity violations', lie in the realm of subatomic physics. Such tiny differences would not show up in these biochemical experiments, says Benner. (He is also interested in making DNA that can avoid unwanted degradation by natural enzymes or viruses, but rather than using mirror-DNA, he has created artificial DNA with non-natural building blocks.)

Church's ultimate goal, to make a mirror-image cell, faces enormous challenges. In nature, RNA is translated into proteins by the ribosome, a complex molecular machine. "Reconstructing a mirror-image of the ribosome would be a daunting task," says Zhu. Instead, Church is trying to mutate a normal ribosome so that it can handle mirror-RNA.

Church says that it is anyone's guess as to which approach might pay off. But he notes that a growing number of researchers are working on looking-glass versions of biochemical processes. "For a while it was a non-field," says Church. "But now it seems very vibrant." ■

BIOTECHNOLOGY

Bankruptcy of nanomedicine firm worries drug developers

Financial troubles of leading biotech firm highlight challenges of making innovative drugs.

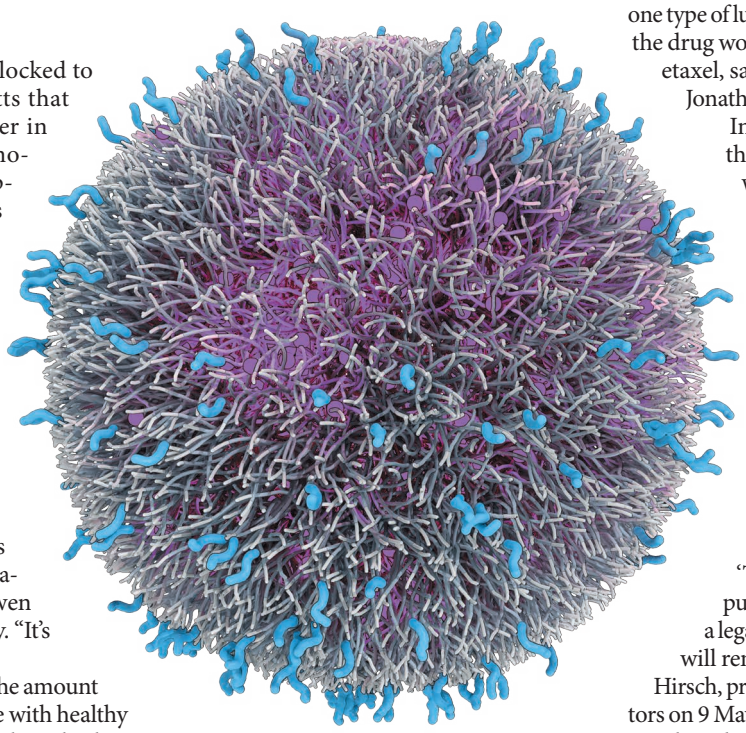
BY HEIDI LEDFORD

Not long ago, investors flocked to a firm in Massachusetts that was hailed as the leader in a wave of next-generation nanotechnology companies developing ways to ferry cancer drugs to tumours. But on 2 May, the company — BIND Therapeutics — declared bankruptcy.

Researchers in the field of nanomedicine are waiting anxiously to see whether the Cambridge-based firm will pull through its financial crisis — and whether its troubles will taint the swiftly evolving field of nanoparticle drug delivery. "It's been a rapid rise and fall," says Eric Schmidt, a biotechnology analyst at the investment bank Cowen and Company in New York City. "It's all unravelled pretty quickly."

Because nanoparticles lessen the amount of contact that cancer drugs have with healthy tissue, they offer a chance to deliver higher doses with fewer side effects. In 1995, the US Food and Drug Administration approved the first such treatment, Doxil, which packages a chemotherapy drug called doxorubicin in a lipid nanoparticle. The particles are too large to escape from normal blood vessels — and so are less toxic to the heart than naked doxorubicin — but they can seep out of the leaky blood vessels often found in tumours.

BIND's nanoparticles were designed to target tumours more precisely than liposome particles can. The company's lead product, BIND-014, involves a polymer particle coated with a molecule that steers the particle to a protein found in many tumours. The particle releases the chemotherapy drug it carries, called docetaxel, inside the tumour.



COURTESY BIND THERAPEUTICS

BIND Therapeutics' nanoparticle is coated in molecules that target it to tumours.

Early tests in animals and small clinical trials showed that the approach was safer than docetaxel alone — and fuelled BIND's US$70.5-million initial public offering in 2013. But later clinical trials disappointed. BIND-014 failed against cervical and head-and-neck cancers. Although it was somewhat effective against one type of lung cancer, it was not clear whether the drug worked any better than regular docetaxel, says BIND's chief scientific officer Jonathan Yingling.

In April, the company announced that it would cut back on its work with BIND-014, and Yingling says that the firm will now explore new targets. It cut the number of employees by 38% and aims to trim its expenses to $6 million per quarter — a dramatic decrease for a company that spent $11 million on research and development alone in the first quarter of 2016.

After one of its creditors demanded that BIND repay a loan ahead of schedule, the company filed for bankruptcy (see 'Troubled times'). It plans to dispute the need for early repayment at a legal hearing on 18 May. "BIND is and will remain open for business," Andrew Hirsch, president of the company, told investors on 9 May.

Schmidt says that BIND remains at the technological forefront of nanoparticle drug delivery, but waited too long to move away from BIND-014. By then, the investor enthusiasm for biotechnology that had driven BIND's initial public offering had cooled. "People are not interested in funding technology right now," Schmidt says. "They're interested in funding later-stage projects. And the one at this company didn't have what it takes."

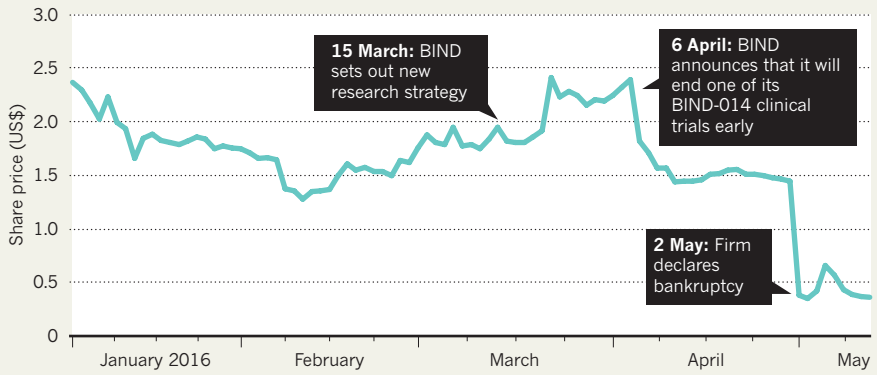
In the time since BIND-014 was developed, researchers have also realized that differences between tumours — such as size, density and leakiness of the blood vessels that lace through

© 2016 Macmillan Publishers Limited. All rights reserved.

SOURCE: YAHOO FINANCE

TROUBLED TIMES

BIND Therapeutics raised US$70.5 million in an initial public offering of stock in September 2013. But the company's stock price has fallen in response to its recent financial woes.



them — can affect how well nanoparticles penetrate them, says Warren Chan, a biomedical engineer at the University of Toronto in Canada. "You should eventually be able to personalize the nanoparticles to the need," he says. "It's just that we're not even close to there yet."

PERSONAL APPROACH

Omid Farokhzad, who studies nanomedicine at Brigham and Women's Hospital in Boston, Massachusetts, and is a co-founder of BIND, thinks that companies should determine whether a nanoparticle would penetrate a patient's tumour before administering the therapy. Farokhzad points to Merrimack Pharmaceuticals, also in Cambridge, which used an imaging agent called ferumoxytol to assess the effects of its liposome-encased chemotherapy in a clinical trial last year. Early results suggested that tumours that took up ferumoxytol were more likely to respond to the nanoparticle-encased drug.

Meanwhile, a third generation of nanoparticles is in the works: particles that shuttle larger molecules such as RNAs, says Pieter Cullis, who studies nanomedicine at the University of British Columbia in Vancouver, Canada. Cullis has been working with Alnylam, a biotechnology firm also in Cambridge, to develop nanoparticles that deliver therapeutic RNA molecules to the liver. The excitement over gene-editing techniques such as CRISPR–Cas9, which could one day be used to correct disease-causing genes, is also fuelling interest in using nanoparticles to shuttle in the RNAs and proteins needed for the method, he says.

But some are worried that BIND's struggles could frighten investors away from the field. Mark Davis, a chemical engineer at the California Institute of Technology in Pasadena, is quick to note that other nanoparticles are showing promise. Last month, Celator Pharmaceuticals of Ewing, New Jersey, announced that its liposome-wrapped chemotherapies are effective against a form of leukaemia. And Cerulean Pharma of Waltham, Massachusetts — a company that Davis co-founded — is making a polymer-based, targeted nanoparticle that has shown promising results in early trials when combined with other chemotherapies.

"I personally would be shocked if 20 years from now, we're not seeing a lot of nanotechnologies helping a lot of people," says Robert Langer, a bioengineer at the Massachusetts Institute of Technology and a co-founder of BIND. "It will happen." ■

© 2016 Macmillan Publishers Limited. All rights reserved.



JASON EDWARDS/GETTY

METAL MIXOLOGY

Stronger, tougher, stretchier: with a simple new recipe, metallurgists are creating a generation of alloys with remarkable properties.

BY XIAOZHI LIM

High-entropy alloys could be used in hotter-running jet-engine blades.

At first glance, the machine seems to be building a miniature cityscape. A ring of nozzles fires four jets of powdered metal into a downward-pointed laser beam, which fuses the colliding grains in a bright orange glow. The mixed grains then solidify on the growing tip of a small pillar of metal alloy. Once the pillar is 1–2 centimetres high, the platform that holds it shifts to the side, and the machine starts to build another one right next door. The result looks like a forest of toy skyscrapers.

In reality, these towers, generated at the Ames Laboratory in Iowa, reflect a major shift in how researchers think about alloys. The standard recipe — used for technologies ranging from ancient swords and arrowheads to modern jet-engine turbines — is to take a useful metal and mix in a pinch of this or a touch of that to improve its properties. One classic example is the addition of carbon to iron to make steel.

But the machine at Ames is making experimental samples of 'high-entropy' alloys, which consist of four, five or more elements mixed together in roughly equal ratios. This deceptively simple recipe can yield alloys that are lighter and stronger than their conventional counterparts, while being much more resistant to corrosion, radiation or severe wear. Eventually, researchers hope, this approach could even produce alloys that have magnetic or electrical properties never seen before, leading to whole new generations of technology.

"We have almost explored everything for traditional alloys," says Yong Zhang, a materials scientist with the State Key Laboratory for Advanced Metals and Materials at the University of Science and Technology Beijing. "For high-entropy alloys, the science is very new," he says — so new that no such alloy has yet made the leap from lab to market. But some researchers are working to make that happen, eyeing potential applications that range from high-temperature furnace linings to ultralightweight aerospace materials. And the field has attracted funding from research agencies in China, Europe, the United States and elsewhere.

"We're not talking about a narrow class of materials, but an extremely broad philosophy on how to combine elements," says Daniel Miracle, a materials scientist at the Air Force Research Laboratory at the Wright-Patterson Air Force Base in Ohio. "The opportunity to find something new and exciting is very high." Last year, he and his colleagues estimated[1] that almost 313,560 different alloys can be made by combining exactly equal proportions of 3, 4, 5 or 6 metallic elements from a set of just 26. More possibilities can come from varying the proportions or expanding the choice of elements.

But not every combination is a winner, says Easo George, a materials engineer at Ruhr University Bochum in Germany. Scientists are still learning what works and what doesn't. Still, he says, "the space available for exploration is really huge, and we have only looked at a small portion of the Universe".

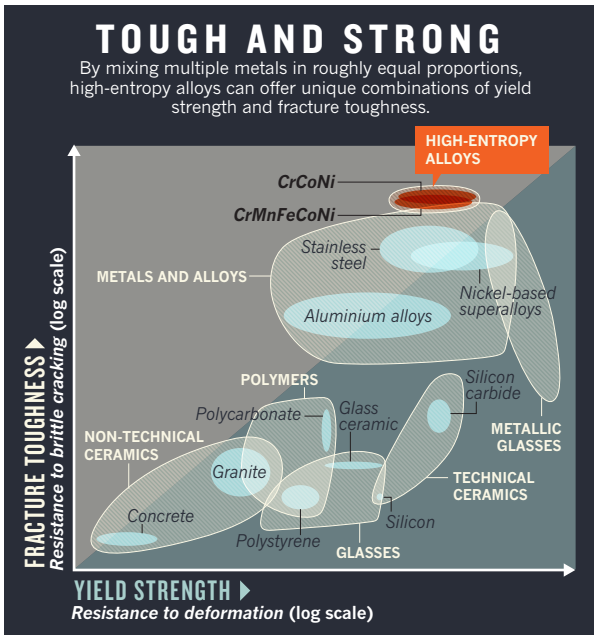
The idea for high-entropy alloys first struck metallurgist Jien-Wei Yeh in 1995, while he was driving through the Taiwanese countryside. The physics of conventional alloys was already well understood, says Yeh, who works at National Tsing Hua University in Hsinchu, Taiwan. At the atomic level, pure metals have a regular crystal structure that stacks layer upon layer of identical atoms. Often, these layers slip past each other easily, which makes the metal too soft to be useful. That is

© 2016 Macmillan Publishers Limited. All rights reserved.

why pure gold is rarely used in jewellery: it cannot survive much wear and tear. But if a metalsmith mixes in an element with a different atomic size, the interloper will randomly disrupt the layers and reduce their tendency to slip, which creates a much harder alloy. The correct choice of compounds can enable metallurgists to tailor other properties as well, such as corrosion resistance or melting point.

But Yeh was also well aware of the potential complications. If too much of the alloying element were added, for example, its atoms could stop falling randomly among the layers, and instead might start to alternate with the primary metal atoms in a more regular pattern, producing a compound that is weak and brittle.

And that gave him an idea, says Yeh: instead of starting with one primary material and mixing in small quantities of one or two elements, why not stir together similar quantities of four or five elements — or even more? The number of possible ways for the different atoms to arrange themselves would expand dramatically, leading to a tendency towards disorder, or 'high entropy', that would overwhelm any bias favouring a regular crystal-lattice structure. Because each of the randomly mixed elements would be a different size, the atoms would become lodged into place and less able to slide past each other, creating a very hard material (see 'Tough and strong').



SOURCE: ADAPTED FROM FIG. 4 IN GLUDOVATZ, B. *ET AL. NATURE COMMUN.* **7,** 10602 (2016).

It was a strange idea at the time, and Yeh gave it a low priority even in his own laboratory; it wasn't until 2004 that his research group first reported success at mixing five to ten elements together[2], producing alloys that were considerably harder than stainless steel. Another group independently announced similar findings around the same time[3].

SPOILT FOR CHOICE

The field began to move rapidly after that. In 2009, Zhang described[4] an alloy composed of cobalt, chromium, copper, iron, nickel and aluminium that was more than 14 times stronger than pure aluminium, but had nearly 3 times the ductility — a measure of a metal's ability to stretch without breaking. In 2011, Yeh devised[5] a cobalt, chromium, iron, nickel, aluminium and titanium alloy that is twice as resistant to damage from friction as conventional wear-resistant steels. And in 2014, George and his team concocted[6] a cobalt, chromium, iron, manganese and nickel alloy that can be cooled below liquid-nitrogen temperatures without becoming brittle. The material could be useful for cryogenic vessels, natural-gas pipelines and other low-temperature applications, such as spacecraft.

Often, however, the best thing about high-entropy alloys — the massive number of possibilities — can also be researchers' biggest challenge. With more than 80 metallic elements in the periodic table, says Miracle, "there are way too many alloys to test, and not enough time". For his own work on high-entropy alloys for aircraft engines and aeroplane frames, he is searching for materials that are lighter, more resistant to corrosion and better able to maintain their strength at high temperatures than anything currently available. To cope with the abundance of choices, Miracle is focusing on elements such as niobium, tantalum and chromium, which have high melting points to begin with.

Another strategy is to try to replicate the characteristics of alloys that are known to work well. For example, some steels are not just a random mix of atoms, but instead contain small nodules of compounds that form when the steels cool rapidly. Although such a composite structure is less stable than a random mixture, it confers high ductility to the steels. Cem Tasan, a metallurgist at the Massachusetts Institute of Technology in Cambridge, has used that knowledge to blend iron, manganese, cobalt and chromium into a high-entropy alloy[7] that is simultaneously extremely hard and highly ductile — traits that once seemed completely incompatible. "It doesn't make sense to abandon all that we know," he says.

The mini-skyscrapers at Ames represent another, more systematic approach. The machine can build up to 30 pillars in less than an hour with a slightly different mix of raw materials in each, so that researchers can test the properties of many alloys quickly. Matthew Kramer, a materials scientist at Ames, leads a project to find high-entropy alloys that can withstand high temperatures and resist corrosion, which could help power plants to operate at higher temperatures and become more efficient.

Aiding his team is Duane Johnson, a theorist at Ames who, in 1995, developed an algorithm to predict the properties of conventional alloys before they are made[8]. In 2015, he expanded the code to work for high-entropy alloys[9]. Johnson's algorithm assesses how much one element is attracted to or repelled by another, and then using that information to predict whether a mixture of elements will form a compound, a solid solution or a mixture of both. That enables Kramer's team to identify which alloys might be worth investigating. The experimental results are then fed back into the algorithm to validate and improve the code.

There are a number of obstacles to overcome to move the high-entropy-alloy field forward. So far, the emphasis has been on improving structural properties, such as strength. But there has been much less work on developing alloys with specific 'functional' properties, including conductivity or response to a magnetic field — a development that would enable applications in areas such as refrigeration and electronics.

Nonetheless, there are many possibilities still to be explored, especially as researchers start to extend the concept well beyond its original definition. Zhang, for example, is mixing metals with elements such as carbon, nitrogen and silicon in an effort to develop new high-temperature ceramics for use in solar-power applications.

And some, including Tasan and Yeh, have begun to experiment with alloys that contain elements mixed in high, but unequal proportions. Their preliminary findings show that many of these still have all the properties that make high-entropy alloys desirable in the first place. For example, Yeh has prepared a range of hard materials that consist of 50% nitrogen, carbon or oxygen, along with a mix of other elements, such as aluminium, silicon or titanium, that can withstand scratching. They could be used as long-lasting coatings for machine parts and cutting tools. "We now have this rich, rich field for exploration," says George. ■

XiaoZhi Lim *is a freelance writer based in Singapore.*

1. Senkov, O. N., Miller, J. D., Miracle, D. B. & Woodward, C. *Nature Commun.* **6,** 6529 (2015).
2. Yeh, J.-W. *et al. Adv. Eng. Mater.* **6,** 299–303 (2004).
3. Cantor, B., Chang, I. T. H., Knight, P. & Vincent, A. J. B. *Mater. Sci. Eng. A* **375–377,** 213–218 (2004).
4. Wang, F., Zhang, Y., Chen, G. & Davies, H. A. *Int. J. Mod. Phys. B* **23,** 1254–1259 (2009).
5. Chuang, M.-H., Tsai, M.-H., Wang, W.-R., Lin, S.-J. & Yeh, J.-W. *Acta Mater.* **59,** 6308–6317 (2011).
6. Gludovatz, B. *et al. Science* **345,** 1153–1158 (2014).
7. Li, Z., Pradeep, K. G., Deng, Y., Raabe, D. & Tasan, C. C. *Nature* http://dx.doi.org/10.1038/nature17981 (2016).
8. Althoff, J.D., Johnson, D.D. & F. J. Pinski *Phys. Rev. Lett.* **74,** 138 (1995).
9. Singh, P., Smirnov, A. V. & Johnson, D. D. *Phys. Rev. B* **91,** 224204 (2015).

© 2016 Macmillan Publishers Limited. All rights reserved.



FRUGAL FARMING

Old-fashioned breeding techniques are bearing more fruit than genetic engineering in developing self-sufficient super plants.

BY NATASHA GILBERT

Jonathan Lynch likes to look beneath the surface. In his quest to breed better crops, the plant physiologist spends a lot of time digging up roots to work out what makes some varieties extremely good at extracting nutrients from the ground. Lynch wants to use this knowledge to develop plants with extra-efficient roots — crops that grow well in the nutrient-starved soils of the developing world. These plants could also reduce the use of fertilizers in richer nations.

Last year, Lynch's forays into the dirt paid off. He and his team at Pennsylvania State University in University Park reported[1] that they had produced a variety of common bean, or string bean (*Phaseolus vulgaris*), with a combination of root traits that allows it to take up phosphorus from the soil with improved efficiency. In experimental plots, the plants produced three times the bean yield of typical varieties.

That result has raised hopes in Africa, where common beans are one of the most important sources of protein for poor people. Researchers in Mozambique are testing how Lynch's beans perform in the country's ecological zones, and they expect to win regulatory approval to bring the crop to market by next year.

Lynch's beans are among the first successful attempts in a global race to develop crops that grow well in soils depleted of nutrients. "Low availability of nitrogen, phosphorus and water are the main limitations of plant growth on Earth. We desperately need this technology," says Lynch.

His work stands out because he has taken an old-school approach. He is leading a renaissance in some conventional crop-breeding techniques that rely on laboriously examining plants' physical characteristics and then selecting for desirable traits, such as growth or the length of fine roots.

And surprisingly, this approach seems to

ERIC KRUSZEWSKI/NGC

© 2016 Macmillan Publishers Limited. All rights reserved.

Farmers in the developing world would benefit from nutrient-efficient crops.

be outpacing the high-tech route.

Big corporations such as DuPont Pioneer in Johnston, Iowa, have spent more than a decade developing improved crops through genetic engineering, and some companies say that their transgenic varieties look promising in field trials. But there are still no fertilizer-frugal transgenic crops on the market, and several agricultural organizations around the globe are reviewing their biotechnology initiatives in this area.

Plant biologist Allen Good of the University of Alberta in Edmonton, Canada, spent years working with companies to develop genetically modified (GM) crops that require little fertilizer, but he says that this approach has not been as fruitful as conventional techniques. The problem is that there are so many genes involved in nutrient uptake and use — and environmental variations alter how they are expressed.

"Nutrient efficiency was supposed to be one of those traits with broad applicability that could make companies lots of money. But they haven't developed the way we thought," says Good.

Despite the scientific and breeding challenges, some researchers say that all strategies must be explored to develop crops that are less nutrient needy. With the global population heading towards 10 billion people by 2050, frugal crops could be essential to feed the planet. "There is a huge worldwide potential for these traits to help increase food production and sustainable development," says Matin Qaim, an agricultural economist at the University of Göttingen in Germany.

MAGIC BEANS

Lynch says that he spent years "groping around in the dark" before he worked out which root characteristics help *P. vulgaris* to grow in low-phosphorus soils. Screening for root traits is a messy and difficult business, so it is unpopular among researchers, he says. That's why most previous research had focused on crop shoots rather than their roots. His team dug up plants to decipher which characteristics were beneficial, taking great care to avoid damaging the fragile root structures. The researchers didn't even know what traits they were looking for in the tangled masses, which could contain thousands of roots per plant.

The key to success is having an open mind and just seeing what the root looks like, says Lynch. "I tell my students they should drop acid before they go to the field, and just look at the plants and let them tell you what they are doing."

One feature that interested Lynch and his team was root depth, because they knew that phosphorus tends to stick to the upper layers of soil. The researchers reported[1] last year that beans with basal roots — those close to the base of a plant — that grow at a shallow angle produced 58% more above-ground biomass than beans with steeper basal roots. The team also discovered that plants with long root hairs took up phosphorus better and improved the bean yield by 89%, relative to the variety with shorter root hairs. When the researchers knitted both the root traits together, they saw much bigger gains. In field trials in Mozambique, beans with shallow roots and long root hairs gave yields of 1,500 kilograms per hectare in soils depleted in phosphorus, compared with local varieties that give just 500 kilograms per hectare.

The beans could be a boon for African farmers who struggle with depleted soils. According to the Food and Agriculture Organization of the United Nations, soils in sub-Saharan Africa lose around 26 kilograms of nitrogen and 3 kilograms of phosphorus per hectare each year through farming and erosion. Because many farmers cannot afford to replenish these nutrients with fertilizers, this stunts yields to about one-tenth of those in developed countries.

The next big research challenge is to improve the beans' tolerance to drought while maintaining their ability to thrive in low-phosphorus soils. This could be tricky because crops under water stress need deep roots — the opposite of what they need to absorb phosphorus, says Magalhaes Miguel, a plant physiologist at the Agricultural Research Institute of Mozambique in Chimoio who collaborates with Lynch. Miguel is testing the beans in different locations around Mozambique to win approval for seed companies to produce and sell the crops.

Lynch's phosphorus discoveries have already made the jump into farmers' fields in China. His colleagues there have shown[2] that some of the same root traits that boost common beans also improve growth in soya beans (*Glycine max*). The regulatory system in China is faster at approving new crop varieties than that in Mozambique, so the improved soya beans are already in use and have been planted on 1 million hectares so far.

Lynch has also had some success in improving crops' uptake of nitrogen, which tends to accumulate deeper in the soil. The ability to grow deep roots is influenced by several factors, including the characteristics of a plant's lateral roots — those that grow out of the main framework of roots. Working in maize, or corn (*Zea mays*), Lynch found that plants perform better under low-nitrogen conditions[3] and water stress[4] when they have just a few long lateral roots rather than many short ones. This arrangement requires less energy to grow, so plants can invest more resources in above-ground vegetation and developing deep roots that can suck up water and nitrogen.

Tests in US fields starved of nitrogen showed[3] that maize with few long lateral roots produced yields more than 30% higher than maize with many short lateral roots. The long-lateral-root variety performed even better in a test of drought-like conditions, where it showed a 144% increase in yield[4].

Other research teams are taking different approaches to breeding fertilizer-frugal crops. Over the past six years, a group at the International Maize and Wheat Improvement Centre (CIMMYT), headquartered in Texcoco, Mexico, has had success using conventional breeding to create maize that grows well in the nitrogen-starved soils of Africa.

Field trials show that CIMMYT's maize produces up to 20% higher yields than existing commercial maize varieties in low-nitrogen soils. By the end of 2015, the centre had won regulatory approval for 14 nitrogen-efficient maize varieties across 6 countries, including Malawi and Kenya. The organization has more than 50 other varieties in the works. Biswanath Das, a maize breeder who leads CIMMYT's project in Nairobi, says that the team is still investigating why some varieties are better than others at finding and using nitrogen. "We hope to have some answers in the next two years," says Das.

"THERE IS A HUGE WORLDWIDE POTENTIAL FOR THESE TRAITS TO HELP INCREASE FOOD PRODUCTION."

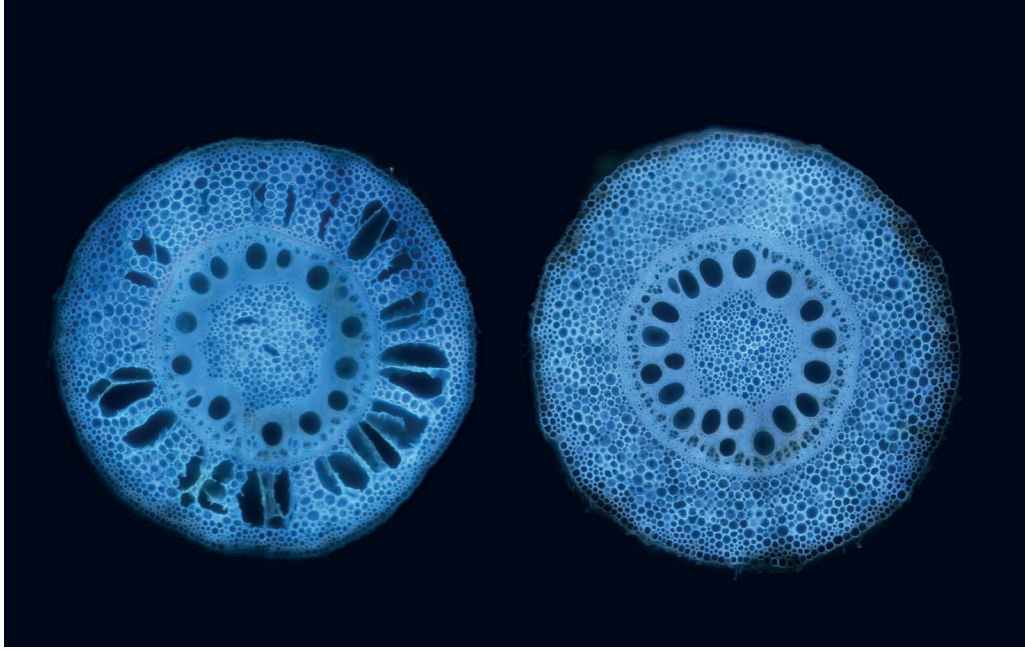
BOOM AND BUST

As Lynch and CIMMYT make big leaps forward, they are overtaking agricultural-biotechnology companies that have invested years of work in tests with GM crops.

Good, who was involved in some of the early experiments, remembers those heady times. In 2001, he walked up and down rows of transgenic and conventional oilseed rape (*Brassica napus*, also known as canola) growing side by side in Californian test fields starved of nitrogen. He marvelled at the contrast between the bigger, greener GM plants, flourishing in the stressful conditions, and their unmodified relatives that were struggling to survive. The GM plants' advantage was a transgene: alanine aminotransferase (*AlaAT*), which came from barley.

"I thought, 'Oh my god, you can see the gene working in the field'," says Good. A few years earlier, he had discovered that *AlaAT* — which has an important role in how plants metabolize nitrogen and incorporate the nutrient into grain — could help crops to thrive

© 2016 Macmillan Publishers Limited. All rights reserved.



JONATHAN LYNCH/PENN STATE

Specially bred maize (left) has roots that transport nutrients better than a conventional plant.

in low-nitrogen soils. The gene is still one of the more advanced targets for transgenic approaches to farmers' over-dependence on fertilizer.

The 2001 trials were run by Arcadia Biosciences, an agricultural-biotechnology company in Davis, California, for which Good was working at the time as a consultant, and to which he had licensed the *AlaAT* discovery. In a collaboration with biotechnology company Monsanto in St Louis, Missouri, separate field studies[5] showed that the transgenic oilseed rape produced 42% more seeds under low-nitrogen conditions than did conventional controls. They also found that the GM crop needed 40% less nitrogen fertilizer to achieve yields equivalent to those of unmodified crops.

Arcadia says that it has successfully put the gene into other crops, including wheat and rice. The company has one of the industry's most robust nitrogen-efficiency programmes, says Roger Salameh, Arcadia's chief operating officer. "We have significant evidence from multiple years of field trials in the two largest food crops in the world — wheat and rice — that we are driving double-digit yield gains," he says.

Despite the positive results, Arcadia has not yet commercialized these or related crops. The company has worked on the nitrogen-efficient wheat with several other companies and governmental agencies, including the Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia's national science agency, but the CSIRO pulled out of that project. A spokeswoman for the organization says that it "did not see results that met its objectives so is no longer actively researching this area".

Arcadia was hoping to have a transgenic variety of nitrogen-efficient wheat on the market this year, but it will miss that target. Salameh says that it is uncertain when the plant will be ready, but the company is now working with Vilmorin and Cie, a seed company in Saint-Beauzire, France, to commercialize the wheat.

FROM THEORY TO PRACTICE

Qaim and some other agricultural analysts are impressed with the results of several transgenic efforts. For example, Arcadia and its research partners say that they have developed a transgenic variety of rice that produced, on average, 30% higher yields than conventional controls during four years of field trials in Colombia under low-nitrogen conditions. However, Qaim says that the crops' performance has yet to be demonstrated in realistic field conditions.

Good has heard that the GM oilseed rape is 'tweaky', meaning that it performs well in some environments but not others. He has lost his initial rapture with it, and says that if he had his time again, he would put the lion's share of his efforts into conventional breeding approaches like those of Lynch and CIMMYT.

Good and the CSIRO are not the only ones worried that the transgenic crops are not yet delivering as hoped. The CIMMYT effort to improve maize yields in low-nitrogen soils kicked off in 2012 with a project to develop a transgenic variety as well as the conventional maize. The transgenic plants are based on technology donated by DuPont Pioneer. They have produced some increases in yields, but not yet enough to warrant the huge effort required to gain regulatory approval, says Das. The transgenic programme is now under review. "The yield benefits are just not that exciting for commercialization yet," he says.

DuPont Pioneer has also seen some delays in its transgenic products targeted to US markets. The company told *Nature* in 2010 that it would have a product available by 2022 (see *Nature* **466,** 548–551; 2010). It now says that the products are still in the proof-of-concept stage, and will not be ready until the middle or late 2020s.

David Spielman, an economist at the International Food Policy Research Institute in Washington DC, says that the delays don't necessarily signal disaster: companies forecast precise timelines to satisfy investors, but the science doesn't always deliver according to corporate plans. The pushed-back timelines do not mean that the scientific advances won't come at some point in the future, he says.

"It is more important to consider if the traits are showing improvement and if public and private sectors are investing in the research with a longer-term view of the path to commercialization," says Spielman.

A key challenge for transgenic efforts is that traits such as nutrient-use efficiency are influenced by a complex web of genes that interact with each other and the environment. "It's very difficult to take one transgene and stick it into a maize variety and expect a big effect," says Das.

DuPont Pioneer and Arcadia are still confident that a single gene will one day deliver the nutrient-efficient goods. "People question if we could find a single gene with a big enough effect, but we are having real success," says Tom Greene, research director at DuPont Pioneer. And Salameh also remains confident about this approach.

Meanwhile, researchers are trying to combine old-fashioned breeding with a variety of high-tech methods. Lynch is now using a computer model that simulates natural selection to work out what combination of root characteristics will boost crops' performance under nutrient stress. He is also identifying genetic markers for useful root traits, which should help to indicate whether plants' offspring have the desired genes and should speed up the breeding processes, he says.

Government regulation could perhaps be the biggest hurdle to both conventional and transgenic frugal crops. Regulatory hurdles in Europe and Africa keep many transgenic technologies from reaching large-scale field testing and commercialization, says Qaim. These hurdles bear part of the blame for the slower-than-expected progress in transgenic frugal crops, he says.

And in some African countries, it can take up to ten years to get new conventional varieties to market because of stringent testing requirements, says Das. Farmers and breeders are also hesitant to take a risk on new varieties.

But if CIMMYT can get the seeds into farmers' fields, the impact on food security and the environment in developing countries could be massive, says Das. "The beauty of it is that the solution is packaged in the seed," he says, so there is no need for extra fertilizer. "Once the farmers have the seed, they have the solution." ■

Natasha Gilbert *is a writer in Washington DC.*

1. Miguel, M. A., Postma, J. A. & Lynch, J. P. *Plant Physiol.* **167,** 1430–1439 (2015).
2. Wang, X., Yan, X. & Liao, H. *Ann. Bot.* **106,** 215–222 (2010).
3. Zhan, A. & Lynch, J. P. *J. Exp. Bot.* **66,** 2055–2065 (2015).
4. Zhan, A., Schneider, H. & Lynch. J. P. *Plant Physiol.* **168,** 1603–1615 (2015).
5. Good, A. G. *et al. Can. J. Bot.* **85,** 252–262 (2007).

© 2016 Macmillan Publishers Limited. All rights reserved.

COMMENT

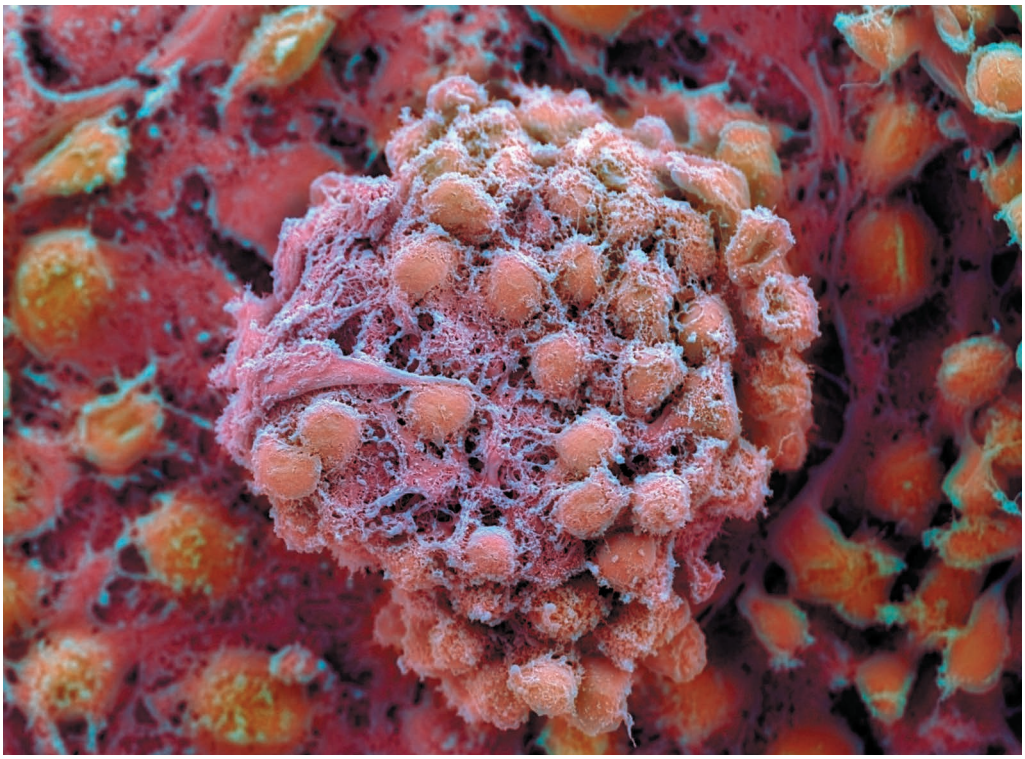
INNOVATION Twelve principles for cross-sector collaboration **p.314**



ECOLOGY Jane Lubchenco enjoys a memoir of science among the sea otters **p.318**

ART Wilderness photography at the improved San Francisco modern-art museum **p.320**

OBITUARY Harry Elderfield, ocean geochemist, remembered **p.322**



MIODRAG STOJKOVIC/SPL

Human embryonic stem cells in culture.

Global standards for stem-cell research

New guidelines from the International Society for Stem Cell Research offer a model for self-regulation in contentious areas, write **Jonathan Kimmelman** and colleagues.

Stem-cell research offers tremendous promise for biomedicine. It also raises vexing ethical and policy challenges. It can involve the destruction, creation and modification of human embryos, and has led to the premature marketing and use of unproven therapies.

On 12 May, in response to scientific progress and evolving ethical concerns, the International Society for Stem Cell Research (ISSCR) issued updated and extended guidelines[1] for work involving the manipulation of stem cells and the translation of that work into medical therapies. The ISSCR is an independent non-profit organization that was established in 2002 to provide a forum for communication and education in the emerging field of stem-cell research and regenerative medicine. The society developed guidelines for embryonic-stem-cell research[2] in 2006 and for clinical translation of stem-cell research[3] in 2008. We represent the working groups that drew up the new guidelines.

The revised ISSCR guidelines provide a model of self-regulation for other potentially contentious research areas. Today's science engages many different actors: researchers, taxpayers, regulators, journals, sponsors, industries and, often, patients. Meanwhile, manuscripts, protocols, tissues and even patients routinely cross national boundaries. In this landscape, different stakeholders need to be confident that their interests will be protected when they collaborate with parties who might have differing views or goals. International guidelines are better positioned than national laws to help ensure protection.

The new ISSCR guidelines span 27 pages. Here we highlight the most dynamic areas for policy, from the introduction of ▶

© 2016 Macmillan Publishers Limited. All rights reserved.

heritable changes into the human genome to the use of sham surgical procedures in the testing of cell-based interventions.

RESEARCH CHALLENGES

Human embryos. In the decade since the ISSCR's previous guidelines were issued, embryo research has entered new arenas. Mitochondrial-replacement techniques (MRT) may soon be used to replace dysfunctional mitochondria in eggs or embryos with those obtained from healthy donors. In the United Kingdom, a pathway for bringing this approach to clinics was approved last year. And a committee convened by the National Academy of Medicine provided recommendations in February to the Food and Drug Administration that would enable clinical testing in the United States.

Mitochondrial diseases result in debilitating physical, developmental and cognitive disabilities. MRT could reduce the chances of women passing mutations associated with these diseases on to their children, but the processes also carry risks that are poorly understood.

"A growing body of research requires the use of fresh human eggs."

More-contentious gene-editing techniques such as CRISPR–Cas9 now enable researchers to modify the nuclear DNA of human sperm and eggs (gametes) and embryos. As with MRT, there are uncertainties about the safety of these techniques. Both MRT and editing the nuclear DNA of human gametes or embryos would introduce potentially heritable alterations into the human genome. Societal consensus is lacking on whether making changes that can be inherited to the genomes of individuals is something that humankind should pursue.

Because of this, the new ISSCR guidelines assert that any attempt to modify the nuclear genome of human embryos for the purpose of human reproduction be prohibited at this time. The revised guidelines do, however, endorse continued laboratory-based research on human embryos and the derivation of stem-cell lines from them.

Just after the first ISSCR guidelines were issued in 2006, scientists reported the derivation of induced pluripotent stem (iPS) cells. These are adult cells that are reprogrammed to an embryonic-like state. Although enormously valuable, iPS cells do not obviate the need for human embryonic stem cells in research[4]. In fact, a better understanding of the different states of pluripotency has renewed biologists' interest in deriving embryonic stem cells with distinct properties. It has also led to a growing recognition that common animal models inadequately recapitulate many aspects of human embryonic development[5].

For research involving human embryos, the revised guidelines assert the need for a specialized embryo-research-oversight (EMRO) process.

Institutions are best positioned to decide what specific mechanism to use to review embryo research. One option may be to repurpose existing embryonic-stem-cell research oversight (ESCRO) committees to take on a broader embryo-research-oversight function. Regardless of process, researchers and reviewers are urged to adhere to certain ethical principles. Among the 'points to consider' listed by the ISSCR are: whether donors of eggs or embryos have provided informed consent; the justification for the study; the number of embryos that will be used; and the quality of the study design.

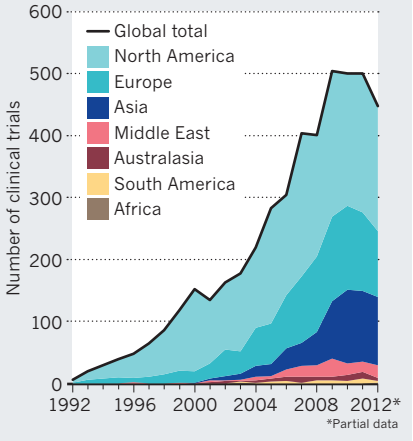
The principles embodied by the revised EMRO process should be applied to the development of MRT, and to investigations of embryo-like structures, which are being increasingly used to model various stages of human development in the lab[6]. Such experiments warrant rigorous EMRO review to eliminate prospects that structures with the potential for integrated human organismal development are kept *in vitro* for anything more than the minimal periods required to address compelling scientific questions.

Human eggs. A growing body of research requires the use of fresh human eggs, whether in mitochondrial replacement, gene editing *in vitro* or nuclear transfer (a form of cloning for research).

Egg donation is invasive and time-consuming: it involves hormone treatment and the retrieval of eggs by needle biopsy. There are also uncertainties about its long-term effects. The practice raises several issues, including how to compensate women for the risks and discomfort but avoid economic exploitation. The new guidelines propose standards. They recommend that when women are paid, the compensation is in line with that received by volunteers in other research[7].

ON THE UP

In recent years, the number of clinical trials involving stem-cell therapies has increased worldwide.



SOURCE: ADAPTED FROM M. D. LI *ET AL. REGEN. MED.* **9**, 27–39 (2016)

Human–animal chimaeras. Various groups are implanting human tissues into the bodies or brains of pigs, non-human primates or rodents. The resulting 'chimaeric' organisms are used to study human organ development and aspects of brain function, and to establish models of human cancer. Such transfers of human tissue raise questions about animal welfare and the limits of permissible chimaera research: it could alter animal cognition or pain perception, or lead to the formation of human gametes in the target animal.

The uncertainty over what is ethically defensible in this area was made apparent last year. The US National Institutes of Health suspended its funding of certain categories of animal–human chimaera research to first evaluate "the ethical issues that should be considered, and the relevant animal welfare concerns"[8].

The ISSCR guidelines offer standards for researchers and reviewers that draw on welfare considerations that are broadly applicable to transgenic animals. They also recommend that certain categories of experiments be prohibited, such as the breeding of non-human animals that might harbour human gametes.

Induced pluripotent stem cells. It is unclear in many national policies whether studies that involve iPS cells should undergo a specialized stem-cell research oversight (SCRO) or an ESCRO review. The ISSCR guidelines recommend that iPS cell work be instead subject to institutional oversight of studies involving human participants, supplemented with stem-cell-relevant informed consent procedures. This would free up SCRO or ethical-review committees to focus on ethically sensitive research activities involving embryos.

CLINICAL CHALLENGES

Irreproducible results and the incomplete reporting of findings from preclinical studies are of particular concern for emerging interventions involving the transfer of living human cells. Decades of research have yielded some general insights about the behaviour of drugs in people. By contrast, the mechanisms underlying potential cellular therapies remain poorly understood for most tissues. And unlike drugs, which are metabolized and excreted from the body, stem cells and their progeny can persist, sometimes for life.

The ISSCR's 2016 guidelines articulate a detailed set of expectations regarding the design, reporting and systematic review of preclinical evidence. For instance, they advocate that the results of all preclinical studies — positive, negative and inconclusive — be reported in peer-reviewed journals and that investigators conduct a systematic

© 2016 Macmillan Publishers Limited. All rights reserved.



SANDY HUFFAKER/GETTY

Researchers store human embryos donated for research in liquid nitrogen.

review to capture all relevant information before initiating a clinical trial (see 'On the up'). They also stipulate that trials begin only after investigators have achieved a high standard of safety and efficacy in relevant preclinical research, as determined by an independent peer-review process.

The guidelines contain strong statements about transparency in clinical research and endorse the prospective registration of all trials in public databases, regardless of phase. They also call for complete and accurate reporting of results in accordance with standards such as those provided by the CONSORT statement — an evidence-based set of recommendations for reporting randomized trials (see www.consort-statement.org).

Despite vigorous admonitions by scientists and regulators against the premature clinical translation of stem-cell therapies[9], numerous providers continue to deliver ill-defined cell preparations to patients outside of trials. In 2014, for instance, leaked documents from an investigation showed that a provider of an unproven cell-based intervention in Italy had no mechanism for screening cells for pathogens, and that sections of its protocol had been lifted from Wikipedia.

Such practices are worrisome. The adverse effects on people's health that may result threaten to set back more painstaking and methodical research programmes by increasing anxiety in ordinary citizens and in regulators about stem-cell-based medicines. The revised guidelines reiterate the condemnation of such practices, and consider the proper testing of new stem-cell interventions in the context of rigorous trials a matter of professional responsibility.

Increasingly, trials are being funded by patient communities, or people are paying to participate. Patient-funding may facilitate some trials that wouldn't otherwise have been pursued — such as those aimed at helping people with rare conditions. But it also allows projects to bypass the long-established mechanisms of peer review and independent oversight that encourage scientific rigour and safety. The ISSCR guidelines stipulate that studies involving paying patients are permissible only when they are subject to independent review mechanisms that can assess scientific rationale, priority and design.

The marketing of unproven interventions has happened in part because of exaggerated claims about stem-cell treatments in the media. Accordingly, the revised guidelines describe ways in which researchers can responsibly communicate with the public, emphasizing the need for balance, clarity and the avoidance of unrealistic optimism.

EVER EVOLVING

The ISSCR's 2016 guidelines are not intended to be the only or last word. They were developed by 25 scientists and ethicists from Asia, Europe, North America and Australia, with review and feedback from more than 100 individuals and organizations, including regulators, funding bodies, journal editors, patient advocates, researchers and members of the public. Some will consider aspects of the guidelines too permissive; others will find parts too restrictive. Moreover, some countries have well-articulated policies that may supersede them. The UK Human Fertilisation and Embryology Authority, for instance, has well-developed guidance on the use of MRT.

Even as we were finalizing the latest guidelines, new ethical challenges were emerging — such as questions about biosafety; whether experimentally generated, self-organizing embryonic tissues should be treated in the same way as human embryos; and whether to revisit the '14-day rule' that limits the culturing of human-embryos *in vitro* to two weeks[6]. The guidelines are intended to be revisited as science and social priorities evolve.

The many limitations of such an aspirational document notwithstanding, we believe that the ISSCR guidelines are well positioned to secure a common ethical basis for stem-cell research.

When widely embraced, international guidelines for professional conduct can be more effective than laws and regulations. The latter are confined to single jurisdictions, can be blunt regulatory instruments and change too slowly to keep pace with cutting-edge research. In highly politicized areas of science, legislated regulatory frameworks are at risk of being revised with every change in government, creating unpredictability that wastes resources and frustrates medical advances.

Guidelines are not binding, but countries, funders, journals and academic institutions can incorporate them into their policies or use them to foster a culture of compliance. Backed by evidence and sound reasoning, they can even provide evidence of professional standards to courts[10]. Ultimately, guidelines give voice to norms and expectations regarding preclinical evidence, trial design and independent review, providing a structured basis for adjudicating disputes.

The ISSCR guidelines continue the tradition of scientists creating professional standards for the responsible conduct of research. They speak most directly to those engaged in stem-cell research but are also relevant to regulators, journal editors, press officers, physicians, funding bodies and patients. Such a global effort to establish research standards offers a model for other contentious research arenas — from artificial intelligence to climate engineering. ■

Jonathan Kimmelman *is associate professor in biomedical ethics, McGill University, Montreal, Canada.* **Insoo Hyun**, **Nissim Benvenisty**, **Timothy Caulfield**, **Helen E. Heslop**, **Charles E. Murry**, **Douglas Sipp**, **Lorenz Studer**, **Jeremy Sugarman** *and* **George Q. Daley.** *On behalf of the steering committee of the International Society for Stem Cell Research's Guidelines Task Force.*
e-mail: jonathan.kimmelman@mcgill.ca; george.daley@childrens.harvard.edu

1. International Society for Stem Cell Research. *Guidelines for Stem Cell Science and Clinical Translation* (ISSCR, 2016); available at http://www.isscr.org/guidelines2016
2. Daley, G. Q. *et al. Science* **315,** 603–604 (2007).
3. Hyun, I. *et al. Cell Stem Cell* **3,** 607–609 (2008).
4. Hyun, I., Hochedlinger, K., Jaenisch, R. & Yamanaka, S. *Cell Stem Cell* **1,** 367–368 (2007).
5. De Los Angeles, A. *et al. Nature* **525,** 469–478 (2015).
6. Hyun, I., Wilkerson, A. & Johnston, J. *et al. Nature* **533,** 169–171 (2016).
7. Haimes, E. *et al. Cell Stem Cell* **12,** 285–291 (2013).
8. US National Institutes of Health. *NIH Research Involving Introduction of Human Pluripotent Cells into Non-Human Vertebrate Animal Pre-Gastrulation Embryos* (NIH, 2015).
9. Taylor, P. L. *et al. Cell Stem Cell* **7,** 43–49 (2010).
10. Campbell, A. & Glass, K. C. *McGill Law J.* **46,** 473–489 (2001).

Full author affiliations and a list of co-signatories accompany this Comment online. Some authors declare competing financial interests. See go.nature.com/pcimjy for details.

© 2016 Macmillan Publishers Limited. All rights reserved.



EASYMILE/ALAIN HERZOG

Driverless buses are being tested by an innovation consortium in Kista, Sweden.

Twelve principles for open innovation 2.0

Evolve governance structures, practices and metrics to accelerate innovation in an era of digital connectivity, writes **Martin Curley**.

A new mode of innovation is emerging that blurs the lines between universities, industry, governments and communities. It exploits disruptive technologies — such as cloud computing, the Internet of Things and big data — to solve societal challenges sustainably and profitably, and more quickly and ably than before. It is called open innovation 2.0 (ref. 1).

The promise is sustainable, intelligent living: innovations drive economic growth and improve quality of life while reducing environmental impact and resource use. For example, a dynamic congestion-charging system can adjust traffic flow and offer incentives to use park-and-ride schemes, guided by real-time traffic levels and air quality. Car-to-car communication could manage traffic to minimize transit times and emissions and eliminate road deaths from collisions. Smart electricity grids lower costs, integrate renewable energies and balance loads. Health-care monitoring enables early interventions, improving life quality and reducing care costs.

Such innovations are being tested in 'living labs' in hundreds of cities. In Dublin, for example, the city council has partnered with my company, the technology firm Intel (of which I am a vice-president), to install a pilot network of sensors to improve flood management by measuring local rain fall and river levels, and detecting blocked drains. Eindhoven in the Netherlands is working with electronics firm Philips and others to develop intelligent street lighting. Communications-technology firm Ericsson, the KTH Royal Institute of Technology, IBM and others are collaborating to test self-driving buses in Kista, Sweden.

Yet many institutions and companies remain unaware of this radical shift. They often confuse invention and innovation. Invention is the creation of a technology or method. Innovation concerns the use of that technology or method to create value. The agile approaches needed for open innovation 2.0 conflict with the 'command and control' organizations of the industrial age (see 'How innovation modes have evolved'). Institutional or societal cultures can inhibit user and citizen involvement. Intellectual-property (IP) models may inhibit collaboration. Government funders can stifle the emergence

© 2016 Macmillan Publishers Limited. All rights reserved.

of ideas by requiring that detailed descriptions of proposed work are specified before research can begin. Measures of success, such as citations, discount innovation and impact. Policymaking lags behind the market place.

The challenge is how to execute and govern the new mode. Innovation is a risky business that has high failure rates — 96% of all innovations do not return their capital cost, and 66% of new products fail within two years. But the potential benefits are vast. Innovation policies should recognize that the linear research-and-development model will be outpaced by a nonlinear, open and collaborative innovation process where the mantra is 'fail fast, scale fast'.

"The mantra is 'fail fast, scale fast'."

Awareness of open innovation 2.0 needs to be raised across industry and society. Here I outline the concept and how it works, and offer a set of 'design patterns' — general solutions to common problems. Adopting these can accelerate the move from conceptual to concrete. The European Union's Open Innovation Strategy and Policy Group, of which I am a chair, is a global leader in distilling its knowledge in this way. The goal is that open innovation 2.0 will become a discipline practised by many rather than an art mastered by few.

DIVERSITY COUNTS

Technical innovation is no longer solely the domain of lone scientists in labs. Scientists in the mid-twentieth century at Bell Labs in Murray Hill, New Jersey, are credited with inventing the transistor, the laser and the Unix computer operating system. Today, a more than one-third of the product launches by consumer-goods firm Procter & Gamble — including the fabric softener Bounce — emanate from ideas that started outside the company.

The term open innovation — where ideas pass between different organizations to create value — was coined by organizational theorist Henry Chesbrough[2] in 2003. Today, the concept is evolving fast. Driven by plummeting communication costs and the ever increasing numbers of connected people and devices, it has never been so easy to exchange information and ideas.

For example, InnoCentive, founded by US drug company Eli Lilly, is a website that matches problems posted by companies with scientists who can potentially solve them. So far, more than 59,000 solutions have been proposed in response to about 2,000 challenges with US$48 million paid to solvers. The GreenTouch consortium, led by telecommunications company Alcatel-Lucent, set a goal of community-sourcing ways to improve the energy efficiency of communication networks by a factor of 1,000 by 2020. It delivered a road map for a factor of 10,000. And the ▶

TWELVE PATTERNS

Keys to collaborative innovation

1 **Purpose.** Efforts and intellects aligned through commitment rather than compliance deliver an impact greater than the sum of their parts. A great example is former US President John F. Kennedy's vision of putting a man on the Moon. Articulating a shared value[4] that can be created is important. A win–win scenario is more sustainable than a win–lose outcome.

2 **Partner.** The 'quadruple helix' of government, industry, academia and citizens joining forces aligns goals, amplifies resources, attenuates risk and accelerates progress. A collaboration between Intel, University College London, Imperial College London and Innovate UK's Future Cities Catapult is working in the Intel Collaborative Research Institute to improve people's well-being in cities, for example to enable reduction of air pollution.

3 **Platform.** An environment for collaboration is a basic requirement[5]. Platforms should be integrated and modular, allowing a plug-and-play approach. They must be open to ensure low barriers to use, catalysing the evolution of a community. Challenges in security, standards, trust and privacy need to be addressed. For example, the Open Connectivity Foundation is securing interoperability for the Internet of Things.

4 **Possibilities.** Returns may not come from a product but from the business model that enabled it, a better process or a new user experience. Strategic tools are available, such as industrial designer Larry Keeley's breakdown of innovations into ten types in four categories: finance, process, offerings and delivery[6].

5 **Plan.** Adoption and scale should be the focus of innovation efforts, not product creation. Around 20% of value is created when an innovation is established; more than 80% comes when it is widely adopted[7]. Focus on the 'four Us': utility (value to the user); usability; user experience; and ubiquity (designing in network effects).

6 **Pyramid.** Enable users to drive innovation[8]. They inspired two-thirds of innovations in semiconductors and printed circuit boards, for example. Lego Ideas encourages children and others to submit product proposals — submitters must get 10,000 supporters for their idea to be reviewed. Successful inventors get 1% of royalties.

7 **Problem.** Most innovations come from a stated need. Ethnographic research with users, customers or the environment can identify problems and support brainstorming of solutions. Create a road map to ensure the shortest path to a solution.

8 **Prototype.** Solutions need to be tested and improved through rapid experimentation with users and citizens. Prototyping shows how applicable a solution is, reduces the risks of failures and can reveal pain points. 'Hackathons', where developers come together to rapidly try things, are increasingly common.

9 **Pilot.** Projects need to be implemented in the real world on small scales first. The Intel Collaborative Research Institute runs research projects in London's parks, neighbourhoods and schools. Barcelona's Laboratori — which involves the quadruple helix — is pioneering open 'living lab' methods in the city to boost culture, knowledge, creativity and innovation.

10 **Product.** Prototypes need to be converted into viable commercial products or services through scaling up and new infrastructure globally. Cloud computing allows even small start-ups to scale with volume, velocity and resilience.

11 **Product service systems.** Organizations need to move from just delivering products to also delivering related services that improve sustainability as well as profitability. Rolls-Royce sells 'power by the hour' — hours of flight time rather than jet engines — enabled by advanced telemetry. The ultimate goal of open innovation 2.0 is a circular or performance economy[9], focused on services and reuse rather than consumption and waste.

12 **Process.** Innovation is a team sport. Organizations, ecosystems and communities should measure, manage and improve their innovation processes to deliver results that are predictable, probable and profitable. Agile methods supported by automation shorten the time from idea to implementation.

© 2016 Macmillan Publishers Limited. All rights reserved.

Innovation Value Institute at Maynooth University in Ireland worked internationally and with competing companies to develop a framework for measuring and improving IT capability, which is now used by hundreds of organizations worldwide.

Companies are opening up their research labs. Philips has converted its research facility in Eindhoven, which had 2,400 employees in 2001, to an open research campus (High Tech Campus Eindhoven) that now houses more than 140 firms and around 10,000 researchers. Breakthrough ideas often emerge at the intersection of disciplines. For example, Keenan, an Irish agricultural-equipment supplier, and telephone company Vodafone have worked with Intel to develop an online service that uses real-time information to provide farmers with nutritional advice for livestock.

Open innovation 2.0 is neither easy nor is it a panacea. It requires courage and energy. But once a critical mass is achieved, innovation can catalyse itself[3]. Just as momentum is the product of mass and velocity, the ecosystem with the most participants and fastest turnover of ideas will be the most successful. Participating organizations must create synergies rather than cancel each other out. High levels of trust and conviction in the shared vision are predictors for eventual success.

A common language helps. Just as architects and engineers can refer to canonical designs when building a bridge, social and technological innovators can improve productivity by following design patterns. These heuristics summarize insights about the innovation process and can be combined (see 'Keys to collaborative innovation'). They shorten learning times and improve the results and pace of innovation.

NEXT STEPS

We have all witnessed how the music and book industries have been transformed by companies such as Apple and Amazon through digitization. Transforming cities, energy grids and health-care systems will be harder. It will need technology breakthroughs, alignment of interests, investment and collaboration across many stakeholders.

Different sectors should explore how disruptive technologies can transform their domain. For example in health care, wirelessly transmitting information from electrocardiograms (a heart-activity test), and using cloud-based analytics, could detect the early onset of conditions such as cardiac fibrillation.

Governments should create innovation strategies that build capacity and focus efforts on nationally important problems. Road maps that set out a sequence of problems to be solved and the innovations required — such as the International Technology Roadmap for Semiconductors — can direct efforts efficiently. Imagine the power of a European or globally agreed road map for smarter cities, more sustainable electrical grids or health-care transformation. The risks of collaboration can be reduced by publishing model contracts for public–private partnerships. Measurement instruments such as Europe's Innovation Union Scoreboard can identify which areas of a country's innovation ecosystem need to be strengthened.

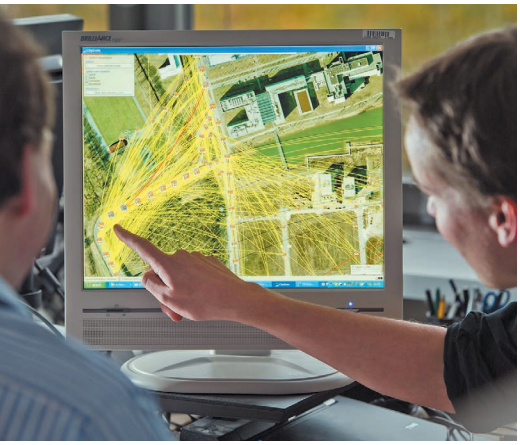
The EU should expand its ambition beyond creating a European Research Area to nurturing a European Innovation Ecosystem. It is promising that a priority for the current Dutch presidency of the EU is ensuring that all European funding instruments focus more on innovation. EU research commissioner Carlos Moedas proposed that the establishment of a European Innovation Council (EIC) would be an important intervention. The EIC could be the steward of an overall European innovation strategy and of societal-challenge road maps.

Governments should encourage the adoption of innovations as well as their creation. They should modulate spending, incentives and policy depending on technology maturity. Immature technologies with high potential need research and pilot projects funded; more-proven prototypes require support for capital deployment cost as well as incentives such as tax credits or loan guarantees. For example, in the United States, incentive payments were used to encourage physicians to adopt electronic health records. Leaders should note the public's increasing desire to be involved in innovation efforts, as exemplified by citizen-science initiatives.

A cultural shift is at the core of open innovation 2.0. The technology is ready — are we? ■

HOW INNOVATION MODES HAVE EVOLVED

Closed innovation	Open innovation	Open innovation 2.0
Dependency	Independency	Interdependency
Subcontracting	Cross-licensing	Cross-fertilization
Solo	Bilateral	Ecosystem
Linear	Linear, leaking	Nonlinear mash-up
Linear subcontracts	Bilateral	Triple or quadruple helix
Planning	Validation, pilots	Experimentation
Control	Management	Orchestration
Win–lose game	Win–win game	Win more–win more
Box thinking	Out of the box	No boxes!
Single entity	Single discipline	Interdisciplinary
Value chain	Value network	Value constellation



Intelligent-lighting sensors can collect useful data about urban activities for city planners.

FRANK VAN BEEK/PHILIPS LIGHTING

Martin Curley *is vice-president at Intel Corporation and director of Intel Labs Europe, based in Dublin, Ireland.*
e-mail: martin.g.curley@intel.com

1. Curley, M. & Salmelin, B. *Open Innovation 2.0 — A New Paradigm* (EU Open Innovation and Strategy Policy Group, 2013).
2. Chesbrough, H. *Open Innovation: The New Growth for Creating and Profiting from Technology* (Harvard Business School Press, 2003).
3. Curley, M. & Formica, P. (eds) *The Experimental Nature of New Venture Creation: Capitalizing on Open Innovation 2.0* (Springer, 2013).
4. Porter, M. E. & Kramer M. R. 'Creating Shared Value' *Harvard Bus. Rev.* (January–February 2011).
5. Gawer, A. (ed) *Platforms, Markets and Innovation* (Edward Elgar, 2011).
6. Keeley, L., Walters, H., Pikkel, R. & Quinn, B. *Ten Types of Innovation: The Discipline of Building Breakthroughs* (Wiley, 2013).
7. Rogers, E. M *Diffusion of Innovations* 5th edn (Simon and Schuster, 2003).
8. Von Hippel, E. *The Sources of Innovation* (Oxford University Press, 1988).
9. Stahel, W. *The Performance Economy* 2nd edn (Palgrave Macmillan, 2010).

CORRECTION

The Comment 'Seven chemical separations to change the world' (D. S. Sholl and R. P. Lively *Nature* **532,** 435–437; 2016) gave the incorrect units for atmospheric distillation. It should have read 230 GW globally.

© 2016 Macmillan Publishers Limited. All rights reserved.



Sea otters are crucial to kelp-forest ecosystems.

FRANCOIS GOHIER/ VWPICS/ ALAMY

ECOLOGY

The sea-otter whisperer

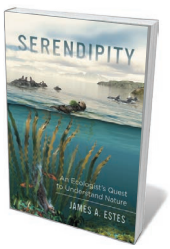
Jane Lubchenco applauds James Estes's chronicle of his 45 years studying the complexities of an apex predator.

The sea otter (*Enhydra lutris*) has had a turbulent history. A cultural icon for tribes living around the coast of the North Pacific Ocean, it became an irresistible target for fur traders who almost drove it to extinction, triggering an international hunting ban in 1911. Then, as the species began to recover in parts of its former range, it became a magnet for tourists, a symbol of hope for marine conservation and the equivalent of Darwin's finches for one scientist: ecologist James Estes. Sea otters are now in decline again in most areas. Estes has relentlessly unpacked insights into the species' history and ecological complexities for more than four decades, in one of the most remote places on Earth.

In *Serendipity*, Estes chronicles that research — in Alaska's Aleutian Islands and beyond — on this relatively small marine mammal, its voracious appetite for crabs, clams, fish and especially sea urchins, and its consequent ability to control the very presence or absence of entire kelp forests and their inhabitants.

In his refreshingly honest and humble narrative, Estes shares the intriguing insights that he gained through careful observation, intentional perturbations of environmental conditions, long-term monitoring, rigorous analysis, willingness to collaborate and openness to new ideas. He acknowledges the often circuitous pathways and serendipitous events that yielded new ideas to test. For example, he relates a pivotal conversation with ecologist Robert Paine, who coined the concept of keystone species, which have a disproportionately large effect on their ecosystems (see *Nature* **493,** 286–289; 2013). That led Estes to pose what would become the defining question of his career — 'What impact do sea otters have on kelp forests?' — rather than his initial query, 'What impact do kelp forests have on sea otters?'

The study of top-down control of an ecosystem by an apex predator and the use of perturbations to test ecological hypotheses were in their infancy in the 1970s, when Estes began his Alaska studies. The Aleutian otters' ongoing recovery from near-extinction provided the perfect mosaic of conditions: in some places, there were many otters; in others, there were few; and in yet others, otters were still absent. Estes' early work demonstrated conclusively that otters in Alaska did control much of the rest of the nearshore community by eating sea urchins, the primary consumers of kelps, and controlling their abundance. Where otters are present, so too are kelp forests and the plethora of other species (fish, crabs, starfish, mussels, snails, seals, seaweeds, birds and more) that depend on the plants for habitat or food. Where otters are absent, urchins thrive, gobble up kelp and keep the seascape barren.



Serendipity: An Ecologist's Quest to Understand Nature
JAMES A. ESTES
University of California Press: 2016.

Estes and his colleagues went on to demonstrate multiple aspects of 'trophic cascades', in which an apex predator's influence ripples through the food web through complex, indirect interactions. Thanks to the team's work, we know that the diets of gulls and bald eagles, the abundance of fish, the growth of mussels and the size and abundance of starfish are all strongly influenced by the presence or absence of otters in Aleutian kelp forests.

Some of the most intriguing ideas that emerged from Estes's studies focus on the surprising and precipitous decline of Aleutian sea otters that began in the early 1990s. Observations of increasing numbers of killer whales, or orcas (*Orcinus orca*), and of killer whales preying on otters, coupled with calculations about otter demography and killer-whale energetics, led to a startling suggestion by Estes and his colleagues: predation by killer whales was driving down otter populations. The team's subsequent hypothesis was even more surprising.

Realizing that something must have changed in the open sea to drive killer whales from their normal oceanic habitats to coastal

© 2016 Macmillan Publishers Limited. All rights reserved.

ones, Estes and his colleagues began to consider broader-scale events in space and time. They knew that after the Second World War, intense industrial whaling in the North Pacific Ocean had reduced whale biomass 'sevenfold to eightfold', up until implementation of the international ban on commercial whaling in the late 1960s and early 1970s. They reasoned that if killer whales had previously fed on great whales, the demise of the latter might have forced killer whales to forage more widely and switch to alternate prey.

Piecing together data from the seas up to 370 kilometres around the Aleutian archipelago, they presented a compelling hypothesis of sequential collapses that spanned four decades. First, great-whale populations plummeted because of whaling. Then, killer whales switched to and decimated first one species, and then another, causing the collapse of populations of harbour seals (*Phoca vitulina*), then Steller sea lions (*Eumetopias jubatus*) and finally sea otters. This 'megafaunal collapse hypothesis', articulated by Estes, Alan Springer and their colleagues, was not received warmly by some scientists who had been studying individual species, but evidence for it continues to mount. Moreover, these ideas raise more intriguing questions: as at least some great whales recover from decimation by whaling, what will killer whales prey on? And what will then happen to seals, sea lions, otters and kelp-forest communities? Let us hope that Estes will continue to provide insights.

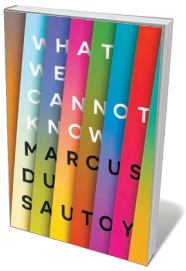
Many of the findings in the book, especially the early ones, are classics of ecology. For non-specialists, Estes's account is a fascinating introduction to key ecological concepts. For the cognoscenti and for students seeking understanding of how science actually works, *Serendipity* is an honest, behind-the-scenes peek at what really happened in the Aleutians and the Pacific more broadly, and what a now-esteemed scientist was thinking and feeling along the way. A rare and delightful insight into timely science. ■

"This relatively small marine mammal has an ability to control the very presence or absence of entire kelp forests."

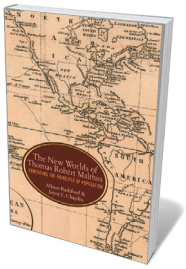
Jane Lubchenco *is a distinguished professor at Oregon State University in Corvallis. A marine ecologist and environmental scientist, she was administrator of the US National Oceanic and Atmospheric Administration from 2009 to 2013, and the first US Science Envoy for the Ocean from 2014 to 2016.*
e-mail: lubchenco@oregonstate.edu

The views expressed are solely those of the author and not necessarily those of the US Department of State.

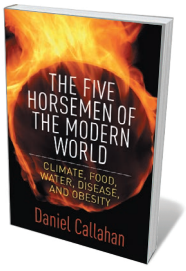
Books in brief



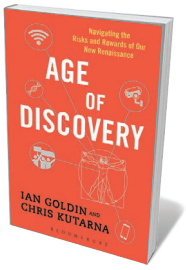
What We Cannot Know: Explorations at the Edge of Knowledge
Marcus du Sautoy FOURTH ESTATE *(2016)*
Science bristles with 'known unknowns', from dark matter to the origins of life. But are some of these conundrums unknowable? In this finely synthesized study, mathematician Marcus du Sautoy explores the edges of our understanding in maths, mind and beyond. Each spiralling investigation begins with an object: casino dice kick-start a foray into probability; a wristwatch propels us into grappling with time. A dazzling journey, vivified by conversations with the likes of neuroscientist Christof Koch on psychophysics and cosmologist Max Tegmark on the mathematical Universe.



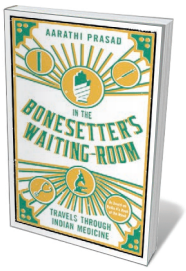
The New Worlds of Thomas Robert Malthus: Rereading the *Principle of Population*
Alison Bashford and Joyce E. Chaplin PRINCETON UNIVERSITY PRESS *(2016)*
"For more than two hundred years, people have loved to hate Thomas Robert Malthus," declare historians Alison Bashford and Joyce Chaplin. In their penetrating reappraisal of the philosopher's *Essay on the Principle of Population*, first published in 1798, they argue that his theory of population growth outstripping agricultural production has been decontextualized. He was less a critic of the English poor than of a swelling European population in his era of colonization, and its impact on people in the New World.



The Five Horsemen of the Modern World: Climate, Food, Water, Disease, and Obesity
Daniel Callahan COLUMBIA UNIVERSITY PRESS *(2016)*
In this trenchant treatise, bioethicist Daniel Callahan investigates why five global issues — water and food shortages, climate change, disease and obesity — remain intractable, despite the billions of dollars thrown at them. Callahan examines cross-cutting problems such as an ageing population, and analyses the role of policy and governance. To ramp up the "intensity level" needed for solutions, he envisions coordinated efforts by grass-roots activists, corporations and national and international agencies. Pragmatic and measured.



Age of Discovery: Navigating the Risks and Rewards of Our New Renaissance
Ian Goldin and Chris Kutarna BLOOMSBURY INFORMATION *(2016)*
The grit and glory of the Renaissance is mirrored in our own age, from looming risks such as pandemics to the transcendent genius needed to overcome them. So aver globalization expert Ian Goldin and political scientist Chris Kutarna in this bold mega-analysis of global education, health, prosperity and technology, then and now. Their parsing of revolutions — from the introduction of the Copernican principle to the digital explosion — is incisive and rich in context and granularity, but not every sweeping statement holds up.



In the Bonesetter's Waiting Room: Travels Through Indian Medicine
Aarathi Prasad PROFILE *(2016)*
Indian medicine is a complex landscape where six traditional systems jostle with Western biomedicine. Science journalist Aarathi Prasad's focused and fascinating journey through the intricacies takes her to Mumbai's sprawling slum Dharavi, where *sanghinis*, or female counsellors, help the hordes of local women subjected to rape and domestic violence. She also visits a clinic that integrates Western labwork with ancient Ayurvedic practice; a project mixing paediatric eye surgery with neuroscience research; and much more. **Barbara Kiser**

© 2016 Macmillan Publishers Limited. All rights reserved.



HENRIK KAM/COURTESY SFMOMA

The San Francisco Museum of Modern Art's extension more than doubles its exhibition space.

ARTS

California on camera

John Gilbey lauds San Francisco's vastly expanded showcase for modern art and photography.

A mid-twentieth-century science-fiction writer's image of the City of the Future: that is how the approach to the San Francisco Museum of Modern Art (SFMOMA) in California can seem. On a clear spring afternoon, the black and white stone oculus of the Mario Botta-designed building emerges from the foothills of the Financial District like the prow of a grounded starship. Now, a new addition more than doubles SFMOMA's exhibition space. The expansion has allowed SFMOMA to remodel public spaces, conservation resources and study areas for a new generation of researchers and visitors.

SFMOMA has had a strong track record in covering the diversity of Bay Area technological innovation and artistic richness. The exhibition 010101: Art in Technological Times, launched in 2001, featured sound installations by musician Brian Eno, as well as video and live digital landscapes — ephemeral work that also brought curatorial challenges. In 2014, the museum showcased geek power at its Art + Data Day, a hackathon that challenged collaborative teams to interrogate the museum collection metadata. Now, with the expansion, SFMOMA boasts the Pritzker Center for Photography, which offers 1,400 square metres to display the museum's renowned photography holdings, including iconic works by nineteenth- and twentieth-century greats such as Carleton Watkins and Ansel Adams. It is now the largest US gallery space dedicated to the technology.

White and multifaceted, the wing's rippled fibreglass-reinforced polymer cladding — designed to echo the Bay Area's famous fogs — has a dynamic look in the changing light. The transdisciplinary approach of its innovative architects Snøhetta integrates architectural, landscape and interior design; their portfolio includes the Bibliotheca Alexandrina in Alexandria, Egypt. The edifice has been costly both financially — at US$305 million — and practically, because the museum building has been closed since June 2013. Until recently, red hoardings enjoined unlucky visitors to "Think Outside the Building" — a reference to its successful touring outreach programme, SFMOMA On The Go.

Inside, the ample room for photography displays gives the museum's 17,000-piece collection real breathing space. SFMOMA has been amassing the holdings, which range from early Daguerreotypes to cutting-edge digital images, since it opened in 1935. This permanent collection is categorized by themes that speak to the unique geographic and cultural position of the city. The local heroes of the Group f/64 collective — Adams, Edward Weston, Imogen Cunningham — are inevitably strongly represented. But the collection of Japanese photography, perhaps the finest outside Japan, demonstrates how San Francisco considers itself to be a Pacific community. Modernist and Surrealist photographers such as Alfred Stieglitz and Man Ray also feature.

Senior curator of photography Sandra Phillips talked to me about the technology's special place in San Francisco — blessed with a photogenic mix of topography, microclimate and infrastructure — and in California, a state endowed with rare geological and ecological richness. Pioneering regional photographers such as Watkins and Adams created a legacy of iconic monochromes. The expansion, Phillips told me, allows for more study as well as display of this deep resource: the Photography Interpretive Center, for instance, enables visitors to explore the lives and work of photographers through video interviews and films.

Phillips is curator of California and the West, an exhibition that traces the portrayal of Californian wilderness and land development in photographs over time. The show, she notes, "will reveal the evolution in our understanding of land use, first as development in mining and logging, then as protection". Along with the work of Robert Adams, who as part of the 1970s New Topographics movement revealed the regional realities of urban encroachment, there will be images by early masters — Dorothea Lange, Minor White and Group f/64. A more contemporary lens on the landscape is provided by Richard Misrach, Larry Sultan, Michael Light and Trevor Paglen.

The expansion offers SFMOMA a way to redefine the role of a modern art museum in a city poised between a vast, sometimes invasive, high-tech culture and an older, perhaps simpler, tradition in a technological medium. And by presenting itself as a public, eclectic resource offering free entry in perpetuity to under-18s, it is setting off beneficial ripples in local culture. Today's young visitors are the digital natives who will shape the future of both technology and the arts. ■

Reopening
San Francisco Museum of Modern Art, California. 14 May.

California and the West
San Francisco Museum of Modern Art, California. 14 May – 5 September.

John Gilbey *is a science and science-fiction writer and photographer.*
Twitter: @John_Gilbey

© 2016 Macmillan Publishers Limited. All rights reserved.

Correspondence

Prevent misuse of eco-compensation

China is piloting a new type of ecological compensation scheme in which developers pay to offset environmental damage. We are concerned that this policy could have the opposite effect.

Ecological compensation by developers in northwest Europe must restore natural habitats and resources to maintain ecological functions and services (see E. Wolanski and M. Elliott *Estuarine Ecohydrology*; Elsevier, 2015). By contrast, ecological restoration is not a priority under China's pilot scheme. Developers may choose to pay government authorities directly rather than opting to restore ecosystems, which is generally more expensive, time-consuming and scientifically demanding.

We are concerned that the payment system could be misused by treating it as 'buying' the right to inflict further damage. We therefore urge the government to develop an ecological compensation policy that is conditional on ecological restoration.

Qinhua Fang *Xiamen University, China.*
Michael Elliott *University of Hull, UK.*
qhfang@xmu.edu.cn

Speed publication of self-corrections

Kudos to Daniele Fanelli for suggesting that authors deserve credit for retracting papers that result from honest errors (*Nature* **531,** 415; 2016). Journal editors and co-authors can be helpful, collaborative and efficient — as they were in the case of two papers I retracted in 2013 (see go.nature.com/jp9meh). However, this is not always the case.

The editorial process can often be lengthy. After waiting for more than two years for the formal resolution of a potential notice of error for another, related paper, I decided to use PubMed Commons to alert the scientific community (see go.nature.com/jzpjfk).

Authors themselves can be uncertain about how to correct the scientific record promptly, even when they suspect errors. They might have moved on to new projects, key contributors may have left the lab, or perhaps they are reluctant to use precious funds to repeat old experiments. Maybe the materials that are needed to repeat the experiments are no longer available, or the authors are slow to reach a consensus on appropriate action.

In my view, these problems could be rectified by standard guidelines for researchers, editors and databases on how to handle self-reported amendments. Terms that are more nuanced than 'correction' or 'retraction' — such as 'notice of concern' or 'error alert' — could be used for situations that are less clear-cut.

Pamela Ronald *University of California, Davis, USA.*
pcronald@ucdavis.edu

EU list should add potential invasives

To be effective, the list of invasive alien species that are targeted for action under the latest European Union regulation needs to focus on preventing new species from entering Europe, and to impede the spread of those that currently have a limited distribution (see also J. Pergl *et al. Nature* **531,** 173; 2016, and F. Santarém *et al. Nature* **532,** 177; 2016).

To be listed under the regulation, species must have been fully risk-assessed. However, existing assessments are mainly for species that are already widespread. All but 2 of the 37 invasive species on the 2015 EU list (see go.nature.com/gigftz) are present in Europe, and 14 are widely dispersed. Of the 95 species that have been identified as potentially invasive (see go.nature.com/avgfb1), only the South American raccoon (*Nasua nasua*) is on the present list.

In our view, post-invasion management is likely to be more successful and cost-effective if the legislation's 'black list' were to focus on potential future invaders as well as on mitigating the impacts of existing invaders.

Maiju Lehtiniemi* *Finnish Environment Institute, Helsinki, Finland.*
maiju.lehtiniemi@ymparisto.fi
**On behalf of 4 correspondents (see go.nature.com/jbvxnz for full list).*

Nicaragua Canal may not benefit shipping

We contend that the benefits of the proposed Nicaragua Canal cannot justify the irreversible damage that it would cause to the environment (see A. Meyer and J. A. Huete-Pérez *Nature* **506,** 287–289; 2014).

With the Panama Canal's widening due to be completed in the next month or so, the argument that the Nicaragua Canal could better accommodate large container ships is no longer valid. Such vessels are of marginal significance to trans-Pacific shipping anyway, owing to insufficient freight demand and port limitations. And as manufacturing bases shift from China to south and southeast Asia, more of Asia's outbound container ships use the Suez Canal to reach eastern US ports.

The Nicaragua Canal would be much longer than the Panama Canal, and so would probably incur higher fees for bulk carriers, tankers and other vessels. On routes that link North America with Asia, large vessels already tend to sail around South Africa's Cape of Good Hope to avoid canal fees.

The Nicaraguan government has also underestimated the safety risks. The country experiences frequent earthquakes, hurricanes and volcanic eruptions — which led the United States to select Panama over Nicaragua for a transoceanic canal in the first place.

Jihong Chen *Shanghai Maritime University, China.*
Xiang Liu *Rutgers University, Piscataway, New Jersey, USA.*
jhchen@shmtu.edu.cn

Clean up the air in Kathmandu

The rebuilding and development of Kathmandu's infrastructure after last year's earthquake in Nepal offer a unique opportunity to improve the city's air quality. According to one ranking, Kathmandu is currently Asia's most polluted city (see go.nature.com/csbt1c).

Kathmandu's air pollution is increasing uncontrollably owing to a lack of environmental legislation in the face of haphazard expansion. The number of vehicles has tripled over the past ten years, and toxic particulates are now "far above" World Health Organization guidelines (A. Gurung and M. L. Bell *J. Expo. Sci. Environ. Epidemiol.* **22,** 235–242; 2012). Unregulated industrial activity, dust from the congested road system and open burning of solid waste all add to the pollution. Moreover, Kathmandu lies in a valley, the topography of which causes temperature inversions that trap this polluted air — making it barely breathable on occasions.

I urge the government to extend its clean-up and restoration beyond buildings, temples and monuments and to draw up stringent environmental regulations. The valley's 3 million inhabitants matter, as well as its tourists.

Keshav Parajuly *University of Southern Denmark, Odense, Denmark.*
kepa@kbm.sdu.dk

CONTRIBUTIONS
See author guidelines for Correspondence submissions at **http://go.nature.com/cmchno**.

© 2016 Macmillan Publishers Limited. All rights reserved.

Harry Elderfield
(1943–2016)

Geochemist who deciphered chemical signatures in the modern and ancient oceans.

Henry 'Harry' Elderfield showed how the distributions of chemicals in seawater and sediments can reveal the ocean's role in historical climate change.

Elderfield, who died on 19 April, was born in 1943, at the height of the Second World War, in North Yorkshire, UK. A few days before his birth, his father, Henry, was reported missing in action, presumed drowned, a loss that may have contributed to Elderfield's draw to the oceans as well as to the past. After obtaining a degree in chemistry from the University of Liverpool, UK, in 1965, he completed a postdoctoral fellowship at Imperial College London, in 1969, and started a lectureship at the University of Leeds, UK, even before completing his PhD.

In 1977, Elderfield headed to the University of Rhode Island in Kingston for a sabbatical year. There he met geochemists Gary Klinkhammer, John Edmond and Wallace Broecker. Discussions with these scientists would have a major influence on the research questions that he later pursued.

Back at Leeds, Elderfield's laboratory technician, Mervyn Greaves, was told to make himself useful in the lab of another faculty member, Chris Hawkesworth. Hawkesworth had established a way to analyse rare-earth metals — a group of 17 elements — in 1-gram rock samples. Elderfield's biggest scientific splash came when he and Greaves applied a modified version of this method to 50-litre samples of seawater and made the first measurements of the tiny concentrations of ten rare earths from samples collected from the North Atlantic. They were present in picomolar amounts (on the order of 10^{-12} moles per kilogram).

Elderfield established how the composition of rare earths in a sample, and the specific properties of the elements, could be used to identify different water masses — deep water bodies from different parts of the ocean that have distinct physical properties. He also predicted — correctly — that these signatures could be extracted from the ocean sediments covering the sea floor and used to track how ocean circulation has changed over time. (The composition of rare earths in the sediment of a particular geological age reflects the composition of the elements in the ocean water at the time that the layer was formed.)

In 1982, Elderfield moved to St Catharine's College at the University of Cambridge, UK. Here he was one of the first to realize how



CAMBRIDGE UNIVERSITY

isotopes that decay radioactively over time could be used to solve marine geochemical puzzles. With Martin Palmer, who had been a student of Elderfield's at Leeds, he developed the strontium isotope curve for seawater for the past 75 million years. Analysing the ratios of strontium isotopes taken from sediments of known age, the pair mapped the changes in the isotopic ratio in the ocean over time.

This isotopic curve, recorded globally in limestone, shows the shifting effects of continental weathering and hydrothermal vents. These sea-floor features spew spectacular plumes of geothermally heated water and chemicals into the ocean, and are the major suppliers of strontium to the ocean. Strontium isotope stratigraphy, a term coined by Elderfield to describe the mapping of strontium isotopes in layers of rock, is now widely used to date sediments. It has also proved useful for climate scientists. Weathering pulls carbon dioxide out of the atmosphere and vents release the gas, so the strontium isotope curve provides a proxy for tectonic processes that are crucial to the carbon cycle.

In 1985, Elderfield's group participated in a cruise that produced the first photographs of a hydrothermal vent called the TAG vent site (P. A. Rona *et al. Nature* **321,** 33–37; 1986). This was the first vent discovered in the Atlantic Ocean or indeed on any slow-spreading mid-ocean ridge. (In subsequent years, Elderfield explored the role of vents in ocean chemistry, as well as fluid flow through the sea floor.)

Around the same time, Elderfield and Edmond, who then was at the Massachusetts Institute of Technology (MIT) in Cambridge, arranged sabbaticals in each others' labs. But as a result of slightly haphazard planning they never overlapped in the same place. At MIT, Elderfield instead spent much of his time with the palaeochemist Edward Boyle, who was then trying to relate the chemistry of the shells of foraminifera, single-celled organisms known as protists, to the chemistry of the seawater in which they lived.

In the shell material, foraminiferal calcite, trace metals of similar size sometimes replace calcium; even trace anions can take the place of the carbonate ion. Elderfield pioneered the exploitation of these 'mineralogical mistakes' to gain insight into past ocean compositions. He established, for instance, that the temperature of the ocean at the time that the foraminifera formed can be inferred from the magnesium–calcium ratio in shells. This ratio provides a way to distinguish between the shifts in oxygen isotopes in the shells caused by changes in ocean temperature from those caused by changes in the size of Earth's polar ice sheets. Ultimately, Elderfield helped to reveal the independent history of changes in global temperature from the waxing and waning of Earth's ice sheets over thousands or millions of years.

Harry was a warm, playful, family man. He had a mischievous sense of humour and a taste for fine art, music, wine and unusual spirits. He also had an incredible knack of teasing sense out of disorder — or in his words, of finding the "really interesting" trends in "unascribable scatter".

I was one of his PhD students in the 1990s. On one tumultuous day I discovered that another group had reported findings similar to those at the heart of my PhD. Harry dissipated my fears with one sentence: "It just shows you are doing exciting science." After a pause and a wry smile, he added, "But let's do it better". ■

Rosalind E. M. Rickaby *is professor of biogeochemistry at the University of Oxford, UK. Harry Elderfield was her PhD supervisor at the University of Cambridge, UK, from 1995 to 1999.*
e-mail: rosalind.rickaby@earth.ox.ac.uk

© 2016 Macmillan Publishers Limited. All rights reserved.

No observable conformational changes in PSII

ARISING FROM C. Kupitz *et al. Nature* **513,** 261–265 (2014); doi:10.1038/nature13453

It is well established[1] that light flashes produce a succession of oxidation events in photosystem II (PSII), driving an initially dark-equilibrated sample (the S_1 state) through a series of states (S_2 and S_3) that represent the accumulation of oxidizing equivalents at the Mn-containing oxygen-evolving complex (OEC), while electrons are transferred to a remote acceptor site. Kupitz *et al.*[2] recently claimed that significant photoinduced conformational changes are observed in PSII using 'diffract before destroy' X-ray crystallography techniques at an X-ray free electron laser (XFEL) source. Although crystallography experiments can potentially give mechanistic insights into PSII-mediated water splitting, the authors' illumination conditions are not suited to move a significant fraction of the system into the S_3 state, and their claim of conformational differences between the S_1 state and the two-flash sample is not supported by their crystallographic data.

Stable charge separation after light absorption is a prerequisite for S-state advancement of the OEC. Owing to the slow acceptor-side kinetics for the final electron transfer step from quinones Q_A to Q_B, a delay between flashes on the order of >100 ms is routinely used for quantitative formation of the higher S states. Although no direct kinetic data are available for the crystals used by Kupitz *et al.*[2], rate constants are available for same-species PSII preparations[3,4]. On the basis of purified protein, the 210 μs flash-spacing used in the XFEL experiment[2] can yield at best 20% of the centres being populated in the S_3 state at the moment of the XFEL probe, 570 μs after the second flash[3–5]. Alternatively, using measured rate constants for intact thylakoids or membrane fragments[6], at most 23% of the centres have reoxidized Q_A 210 μs after the first light flash and are able to undergo stable charge separation after the second flash to form the S_3 state. The reported EPR data (extended data fig. 1f, g in ref. 2) were recorded with a flash-spacing of 1 s and therefore do not indicate successful S-state turnover with the almost 5,000-fold shorter flash-spacing used during the XFEL experiments.

The analysis of the crystallographic electron density is problematic for several reasons. First, no $mF_o - DF_c$ difference maps were reported, although such maps are the most indicative of small changes in the OEC. Second, inspection of the deposited structures (Protein Data Bank (PDB) accession codes 4PBU and 4Q54) revealed that the authors transposed the two PSII monomers in the two-flash state relative to the dark state. As a result, the differences interpreted to arise from flash states could have arisen from the differing crystal environments between the two monomers. Third, the changes in the density surrounding the OEC presented in fig. 3 of ref. 2 are not reproducible using most well-validated methods for map calculation (with or without bias removal; Fig. 1a). Only when generating simulated annealed (SA) omit maps exactly as published[2] using phenix.autobuild (a program that was never intended to reveal fine structural details) were we able to reproduce partially the claimed changes in the region of the OEC, but only for one of the two monomers (Fig. 1b). However, these maps are very sensitive to uncertainties in the atomic model used for phasing, and inspecting a slightly larger area of the SA-omit map produced by the protocol in ref. 2 reveals numerous areas with similar differences between the light and dark states (ellipses in Fig. 2). If these were true structural changes, one would conclude that the entire protein backbone is modified by illumination. However, the most likely interpretation is that the differences are below the significance level of the experiment, given the limited resolution (5.0 and 5.5 Å) of the diffraction data.

In summary, femtosecond X-ray crystallography at XFEL sources promises to provide great insights into complex biological systems such as PSII. However, every care has to be taken to ensure that the correct states are being studied and that the interpretation is consistent with the level of uncertainty in the data. These requirements are not fulfilled in the work by Kupitz *et al.*[2]

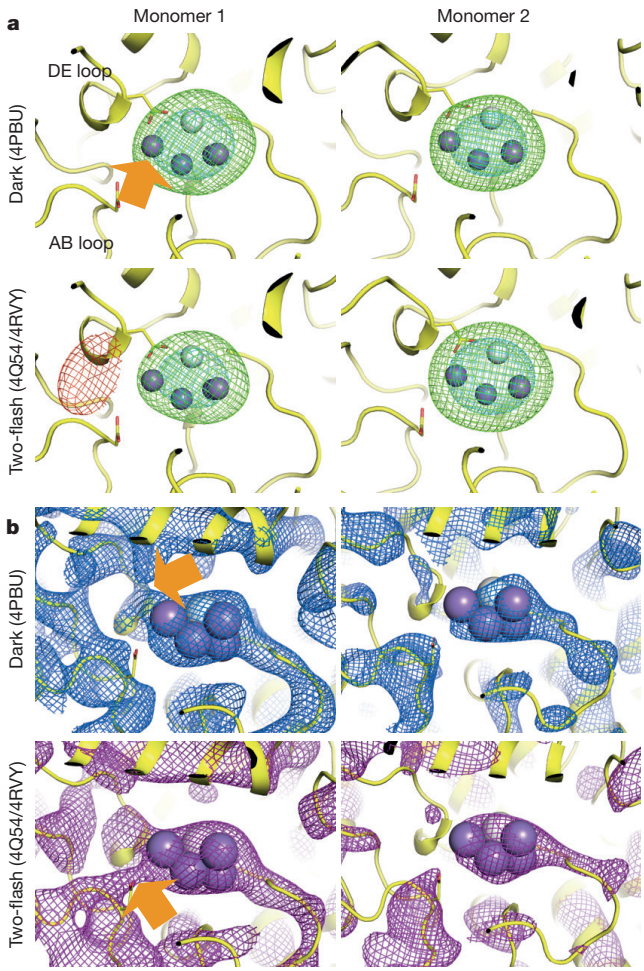


Figure 1 | Reanalysis of the electron density from ref. 2 data. **a**, Simulated annealing $mF_o - DF_c$ omit maps generated by phenix.refine with the OEC set to zero occupancy and surrounding atoms restrained. This is the classical protocol for producing SA-omit maps[11]. Green and red mesh denotes $\pm 3.0\sigma$; cyan mesh denotes $+6.0\sigma$; in the dark (PDB code 4PBU) and two-flash (PDB code 4Q54 with corrected monomer assignment, equivalent to PDB code 4RVY) states. No change of the density around Mn4 (indicated by an orange arrow) is visible when comparing dark and two-flash data. **b**, The phenix.autobuild SA-omit maps recalculated following the protocol of ref. 2, contoured at 1.5σ. Orange arrows indicate the claimed conformational change. No changes are visible in monomer 2. Purple spheres denote Mn; white sphere denotes Ca.

© 2016 Macmillan Publishers Limited. All rights reserved

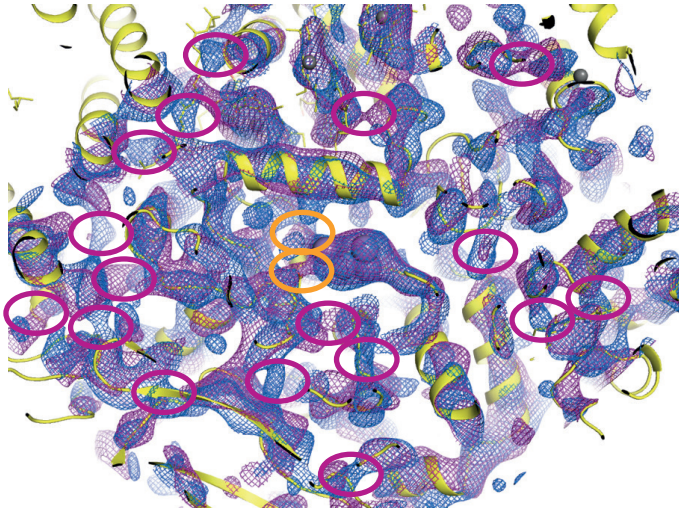


Figure 2 | Electron density variability between the two data sets. View of a larger area around the OEC in monomer 1, showing phenix.autobuild SA-omit maps contoured at 1.5σ as in Fig. 1b (blue denotes 4PBU, purple denotes 4Q54 with corrected monomer assignment as in 4RVY). Orange ellipses denote structural differences claimed by Kupitz *et al.*[2] Red ellipses denote additional map differences throughout the structure, showing that the claimed differences at the OEC are within the noise level of these electron density maps.

Methods

The maximum possible S_3-state population was calculated using $t_{1/2}(S_2 \rightarrow S_3) = 70\,\mu s$, $t_{1/2}(Q_A^- \rightarrow Q_B^-) = 0.54$ ms[4,5], a miss factor of 10%, and 100% quinone occupancy of the Q_B acceptor site. Alternatively, $t_{1/2}\text{fast}(Q_A^- \rightarrow Q_B^-) = 0.2$ ms and $t_{1/2}\text{slow}(Q_A^- \rightarrow Q_B^-) = 1.6$ ms with a ratio of 0.47:0.53, and an overall fraction of centres involved in $Q_A^- \rightarrow Q_B^-$ electron transfer of 0.75 was used[6]. Relative monomer placement between states was analysed with phenix.emma[7], and the molecules placed on the same crystallographic origin using phenix.find_alt_orig_sym_mate[8]. This led to the conclusion that the monomers in the two-flash state coordinate file (PDB code 4Q54) are swapped relative to the dark state file (PDB code 4PBU). After bringing this to the authors' attention, they deposited an updated coordinate file for the two-flash state (PDB code 4RVY) correcting the transposition of monomer labels. SA-omit maps were calculated using the atomic coordinates and experimental data deposited in the PDB for accessions 4PBU and 4Q54 (with correctly swapped monomer assignment as in 4RVY), with OEC atoms set to zero occupancy and using either phenix.refine[9,10] with surrounding atoms tightly restrained[11] (Fig. 1a) or phenix.autobuild[9,10] in the composite_omit mode (Figs 1b and 2).

Nicholas K. Sauter[1], Nathaniel Echols[1], Paul D. Adams[1], Petrus H. Zwart[1], Jan Kern[1,2], Aaron S. Brewster[1], Sergey Koroidov[3]†, Roberto Alonso-Mori[2], Athina Zouni[4], Johannes Messinger[3], Uwe Bergmann[2]†, Junko Yano[1] & Vittal K. Yachandra[1]

[1]Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA.
emails: NKSauter@lbl.gov and VKYachandra@lbl.gov
[2]LCLS, SLAC National Accelerator Laboratory, Menlo Park, California 94025, USA.
[3]Institutionen för Kemi, Kemiskt Biologiskt Centrum, Umeå Universitet, 90187 Umeå, Sweden.
[4]Institut für Biologie, Humboldt-Universität zu Berlin, D-10099 Berlin, Germany.
†Present addresses: Stanford PULSE Institute, SLAC National Accelerator Laboratory, Menlo Park, California 94025, USA (S.K. and U.B.).

Received 17 November 2014; accepted 21 March 2016.

1. Kok, B., Forbush, B. & McGloin, M. Cooperation of charges in photosynthetic O_2 evolution-I. A linear four step mechanism. *Photochem. Photobiol.* **11,** 457–475 (1970).
2. Kupitz, C. *et al.* Serial time-resolved crystallography of photosystem II using a femtosecond X-ray laser. *Nature* **513,** 261–265 (2014).
3. Krivanek, R. *et al.* Spare quinones in the Q_B cavity of crystallized photosystem II from *Thermosynechococcus elongatus*. *Biochim. Biophys. Acta* **1767,** 520–527 (2007).
4. Zimmermann, K. *et al.* Herbicide binding and thermal stability of photosystem II isolated from *Thermosynechococcus elongatus*. *Biochim. Biophys. Acta* **1757,** 106–114 (2006).
5. Noguchi, T. *et al.* Time-resolved infrared detection of the proton and protein dynamics during photosynthetic oxygen evolution. *Biochemistry* **51,** 3205–3214 (2012).
6. de Wijn, R. & van Gorkom, H. J. Kinetics of electron transfer from Q_A to Q_B in photosystem II. *Biochemistry* **40,** 11912–11922 (2001).
7. Grosse-Kunstleve, R. W. & Adams, P. D. On symmetries of substructures. *Acta Crystallogr. D* **59,** 1974–1977 (2003).
8. Oeffner, R. D., Bunkóczi, G. & Read, R. J. phenix.find_alt_orig_sym_mate. *Computational Crystallography Newsletter* **3,** 5–10 (2012).
9. Adams, P. D. *et al. PHENIX*: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
10. Afonine, P. V. *et al.* Towards automated crystallographic structure refinement with *phenix.refine*. *Acta Crystallogr. D* **68,** 352–367 (2012).
11. Hodel, A., Kim, S.-H. & Brünger, A. T. Model bias in macromolecular crystal structures. *Acta Crystallogr. A* **48,** 851–858 (1992).

Author Contributions N.E., P.D.A., P.H.Z., J.K., A.S.B., S.K. and J.M. performed the data analysis. All authors participated in writing the paper.

Competing Financial Interests Declared none.

doi:10.1038/nature17983

© 2016 Macmillan Publishers Limited. All rights reserved

FORUM Neuroscience

Virtual reality explored

Neuroscientists are increasingly using virtual reality to facilitate studies of animal behaviour, but whether behaviour in the virtual world mimics that in real life is a matter for debate. Here, scientists discuss the strengths and limitations of the approach.

THE TOPIC IN BRIEF

- Virtual-reality (VR) systems simulate real-world inputs to one or more of an organism's sensory neural circuits, then measure the subject's actions and apply updates to sensory stimuli in response.
- In most rodent set-ups, the animal receives visual information from an immersive screen that spans its field of vision. The animal's movements control the visual flow, thereby replicating the sensory–motor coupling of the real world.
- Typically, movement is restricted by fixing the rodent's head in position; this allows precise measurements of neural activity to be taken and correlated with motor actions in animals that are awake, rather than anaesthetized (Fig. 1).
- Many researchers think that VR is a valuable tool for studying both navigation and sensory systems.
- However, a body of work[1–3] indicates that the way in which mice navigate in real and virtual worlds is different.

The best of both worlds

MATTHIAS MINDERER
& CHRISTOPHER D. HARVEY

Virtual reality is a valuable tool for understanding neural function because it combines precise experimental control with natural behaviours. It allows experiments that are not possible using real-world approaches. As such, it has increased our understanding of neural processes in subjects ranging from humans to insects.

What are the experimental benefits of VR? First, the technology allows researchers to define explicitly and exhaustively the sensory cues that carry information about the virtual world. In real-world experiments, it is not possible to control all sensory cues. For example, when studying the contribution of visual cues to navigation, confounding information could be provided by unmeasured smells, sounds, textures and vestibular stimuli (internal information about balance and spatial orientation). VR offers the means to add or remove sensory cues to test the contribution of each one to a neural code, and to build up a 'minimal' set of stimuli needed to produce a given behaviour or neural activity pattern.

A second benefit comes from the ability to redefine the laws that link the subject's actions to changes in its world. When an animal explores the real world, it is difficult to disentangle which neural responses are attributable to the animal's actions and which are caused by sensory stimuli, because the two are rigidly linked by the laws of physics. In VR, this link can be modified in informative ways — sensory and motor features can be dissociated by changing the gain or lag between an action and a subsequent update of the virtual environment, or be made independent of one another for brief periods. Sensory and motor variables can therefore be separated while allowing the subject to interact naturally and actively with the sensory world.

Third, VR increases the range of tools available to measure neural activity. Because the subject is usually constrained, techniques can be applied that are either not possible or of poorer quality in freely moving subjects. These include functional magnetic resonance imaging, high-resolution fluorescence imaging and intracellular single-neuron electrophysiology.

Many studies have shown that animals can solve navigational tasks in virtual worlds[4–6]. But the aspects of navigation that can be studied in VR depend on the experimental set-up — for instance, the number of sensory cues simulated, the degree of sensory immersion and how naturally the subject interacts with the virtual world. In VR experiments that provide visual inputs and allow body rotations to trigger vestibular signals, neural activity patterns during navigation are consistent with those measured in real-world experiments[5]. Furthermore, studies[1–3] that remove key sensory inputs such as vestibular stimuli reveal which aspects of navigational neural activity depend on vestibular input and which can be supported by visual cues alone. Therefore, VR can recapitulate neural activity in real environments, and VR experiments can be designed to create informative differences between neural function in real and virtual worlds.

Overall, VR has yielded many insights into sensorimotor integration, decision-making and navigation[6]. But it is important to remember that, like all reductionist approaches, VR requires a trade-off between improved experimental accessibility and consistency with natural processes — the optimum set-up depends on the research question being asked. For instance, in studies of sensorimotor integration, it is crucial to dissociate sensory and motor variables. In navigation studies, convincing simulations are needed to probe the subject's internal model of the physical world. VR must be used judiciously, so that its implementation matches the needs of the question. Of course, this requirement applies to all experimental tools and is not specific to VR.

VR experiments can be designed to create informative differences between neural function in real and virtual worlds.

In summary, we consider VR as bridging the gap between natural behaviour and conventional reductionist approaches; this is a major step forward in the study of complex behaviours in many species. As the community of VR users grows and commercial VR technologies expand, we expect the range of applications for VR to continue to grow, enhancing our understanding of neural function.

Matthias Minderer *and* **Christopher D. Harvey** *are in the Department of Neurobiology, Harvard Medical School, Boston, Massachusetts 02115, USA.*
e-mail: harvey@hms.harvard.edu

© 2016 Macmillan Publishers Limited. All rights reserved.

A world away from reality

FLAVIO DONATO & EDVARD I. MOSER

Technology that involves VR has obvious advantages for studies of simple sensorimotor computations, in which a defined set of inputs, such as those corresponding to an animal's movement, is associated linearly with neural output. However, some pressing concerns are raised when VR technology is used to study higher-order computations such as spatial navigation. Navigation reflects the integration of many sensory inputs. The resulting outputs are not linearly related to sensory perception, but rather express cognitive abstractions.

Goal-driven navigation relies on several cell types in the brain, including place cells (which fire when an animal is in a particular location), grid cells (which fire at periodically spaced positions across the entire environment) and border cells (which fire selectively along local borders)[7,8]. By fixing an animal's head in place, investigators can monitor the activity of these neurons at high resolution while the animal runs between specific locations in virtual space. But do animals navigate in the same way in VR as in real life?

Navigating in the real world is a multisensory process that integrates visual, olfactory and tactile stimuli with vestibular information and information about the activity of moving body parts. But in VR, these elements are often not coordinated, and the animal's sensory experience is largely reduced to a combination of visual inputs and locomotion, which are easy to control. The animal must overcome discrepancies between visual cues that follow movements and cues that are static in VR, such as smell or head direction. Conflicts between movement and sensory inputs might alter the activity of space-encoding neurons to reflect only information coordinated to motion, such as visually changing landmarks and accumulated distance[1,2], at the expense of other cues. This could lead researchers to overestimate the contribution of visual inputs to navigation and, in the most extreme cases, might lead to the loss of computation altogether[2].

A particular concern is whether the loss of vestibular input that accompanies movement restriction affects animals' computation of their position. A continuous mismatch between vestibular and visual inputs might not be detrimental in linear environments. When an animal runs in a straight line, visual inputs are repeatedly and stereotypically paired to the same locomotor information, which may, with continued training, allow the animal to compensate for mismatches. However, such a mismatch might have a greater effect in two-dimensional or 3D VR arenas.

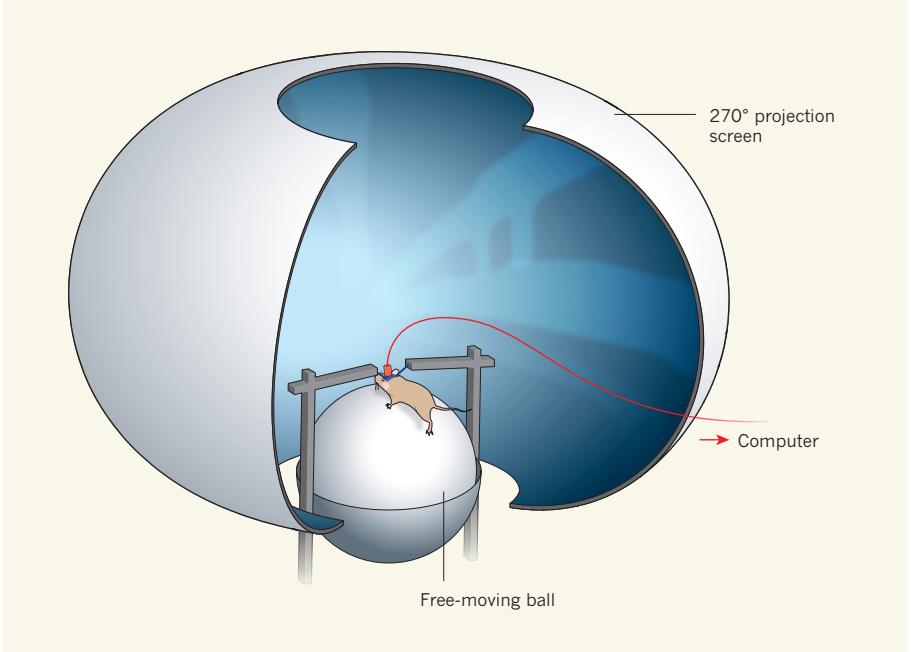


Figure 1 | A mouse explores a virtual world. In most typical virtual-reality experiments, a mouse is head-fixed above a ball. Its legs are free, allowing it to move the ball in all directions. By moving the ball, the mouse navigates around a virtual world that is projected onto a 270° doughnut-shaped screen in front of it. Head fixing enables neural activity to be measured and correlated with the motor actions that drive movement.

Indeed, if movement is unrestrained, the position-coding activity of place and grid cells is similar in 2D VR to that in the real world[5]. In stark contrast to this, position coding is disrupted and a new coding emerges when body movement is restricted, or if the head is fixed[2]. These data cast doubt on whether the way animals interpret 2D or 3D space can ever be understood using VR under conditions of head or body restriction. Strategies that compensate for the loss of synchrony between vestibular information and the animal's behaviour would be a welcome advance.

A particular concern is whether the loss of vestibular input that accompanies movement restriction affects animals' computation of their position.

Finally, are all types of position-coding cell represented in VR-based navigation? It is unclear if and how border, speed and head-direction cells are activated when movement is restricted. Moreover, cells might not fire in the same way in the two worlds. In one analysis[2], 60% of the place cells activated in the real world were silent in VR. Whereas studies typically check that VR-activated cells are represented in real-world sessions, the opposite direction of investigation lags behind — although there are exceptions to this[3].

More than 40 years ago, the neuroscientist John O'Keefe changed our understanding of the physiology of navigation by studying rats freely foraging for food. By allowing the natural sensory–motor interactions required for the formation of an internal representation of space, O'Keefe discovered the first element of the 'cognitive map' — the place cell[9]. VR can extend that ecological approach to higher cognitive functions. But to do so successfully, the technology needs further development and validation. ■

Flavio Donato *and* **Edvard I. Moser** *are at the Kavli Institute for Systems Neuroscience, Norwegian University of Science and Technology, 7491 Trondheim, Norway.*
e-mails: flavio.donato@ntnu.no; edvard.moser@ntnu.no

1. Ravassard, P. *et al. Science* **340,** 1342–1346 (2013).
2. Aghajan, Z. M. *et al. Nature Neurosci.* **18,** 121–128 (2015).
3. Acharya, L. *et al. Cell* **164,** 197–207 (2016).
4. Harvey, C. D., Coen, P. & Tank, D. W. *Nature* **484,** 62–68 (2012).
5. Aronov, D. & Tank, D. W. *Neuron* **84,** 442–456 (2014).
6. Dombeck, D. A. & Reiser, M. B. *Curr. Opin. Neurobiol.* **22,** 3–10 (2012).
7. Moser, E. I., Kropff, E. & Moser, M.-B. *Annu. Rev. Neurosci.* **31,** 69–89 (2008).
8. Rowland, D. C., Roudi, Y., Moser, M.-B. & Moser, E. I. *Annu. Rev. Neurosci.* http://dx.doi.org/10.1146/annurev-neuro-070815-013824 (2016).
9. O'Keefe, J. & Dostrovsky, J. *Brain Res.* **34,** 171–175 (1971).

This article was published online on 11 May 2016.

© 2016 Macmillan Publishers Limited. All rights reserved.

DRUG DISCOVERY

Fighting evolution with chemical synthesis

A synthetic strategy has been developed that provides easy access to structurally diverse analogues of naturally occurring antibiotics, providing a fresh means of attack in the war against drug-resistant bacteria. SEE ARTICLE P.338

MING YAN & PHIL S. BARAN

Pathogenic bacteria have developed strong resistance to various antibiotics as a result of the rapid evolution of the bacteria in response to widespread antibiotic misuse. This problem is exacerbated by the reduction in antibacterial-drug research by large pharmaceutical companies over the past few decades — we are failing to replenish our armamentarium quickly enough to compensate. An apocalyptic scenario looms wherein a drug-resistant 'superbug' becomes so well established that the most trivial infections would be as life-threatening as they were during the pre-antibiotic era. On page 338 of this issue, Seiple *et al.*[1] offer an ingenious strategy to bolster the antibiotic-discovery process: a highly modular and versatile chemical-synthesis approach that provides access to a rich array of antibiotics possessing different molecular topologies and functionalities.

Naturally occurring compounds derived from bacteria or fungi have long been the primary source of antibiotics. These compounds can be produced on a large scale by fermentation, but few inherently possess adequate bioavailability, *in vivo* activity or metabolic stability for use as drugs[2]. Chemical modifications are often necessary to address these issues. The complexity of natural products means that such modifications often involve a long, linear sequence of steps in which chemical groups in the molecules are masked with 'protecting' groups to allow the modifications to be made, and then unmasked. Making variations at certain molecular locations tends to be particularly laborious or impractical, and so synthesizing a diverse collection of antibiotic candidates for drug discovery can be forbiddingly difficult.

Seiple and colleagues show convincingly that chemical synthesis can surmount the challenge of molecular diversification, trumping strategies based on synthetic biology[3] or semi-synthesis (in which compounds from natural sources are used as starting materials). In their approach, the authors constructed complex macrolides — antibiotics that contain a characteristic large ring of atoms known as a macrocycle — through the assembly of eight modular building blocks. These simple modules could each be varied easily to incorporate different structural motifs at specific locations in the macrolide skeleton.

The researchers' synthetic route included seven key coupling reactions, each of which introduced different building blocks into the macrolide (Fig. 1). Such reactions are said to be convergent because they bring together modules that were prepared in separate, parallel reaction sequences. The convergent coupling reactions serve as branch points in the synthesis that allow a diverse range of structural modifications to be easily made. For example, Seiple *et al.* introduced reactive groups such as azides, allyl groups and amides at different positions, all of which could be readily converted into various structural motifs. Such manipulation, when combined with variations in the building blocks, enables an exponential expansion of diversity.

This approach is possible only because the synthetic sequence is extremely well choreographed. Steps such as protecting-group manoeuvres are kept to a minimum; convergent couplings and other complexity-building transformations account for the bulk of the sequence. These key reactions are mechanistically very different from one another (ranging between the addition of an enolate and a transition-metal-catalysed cycloaddition, for those in the know). Yet the authors show that these all work well for modules that carry a variety of groups.

The formation of macrocyclic rings is one of the toughest challenges in macrolide synthesis — many routes to these antibiotics are foiled because attempts to form the ring fail[4]. Conventionally, macrocycle formation must be individually optimized for each reaction substrate by finding a set of protecting groups that allows the transformation to occur. However, Seiple and colleagues report that, in their work, macrolides containing various ring sizes can be made using almost identical ring-formation conditions.

Similarly, other reactions in the synthetic sequence need minimal adaptation to work for different substrates that contain a variety of chemical groups, and deliver consistent yields of products. Such generality is the cornerstone of the authors' approach, because it allows hundreds of analogues of macrolides, bearing highly dissimilar arrangements of attached groups, to be accessed expeditiously with little perturbation to the overall strategy.

Another prerequisite for rapidly generating structural diversity is an abundant supply of diversifiable intermediates; this in turn requires all the chemical transformations in the syntheses to be conducted on gram scales[5]. Seiple *et al.* tackled this challenge using a series of in-house methodologies. The first was a scalable aldol reaction, a type of carbon–carbon bond-forming reaction. This reaction shows exceptional stereoselectivity (it preferentially forms isomers that have specific 3D orientations of atoms) and tolerates the presence of a potentially interfering amino group (NH_2) in one of the reactants[6]. The second was a concise approach[7] to preparing the carbohydrate desosamine and its analogues, which form a key part of another building block. These

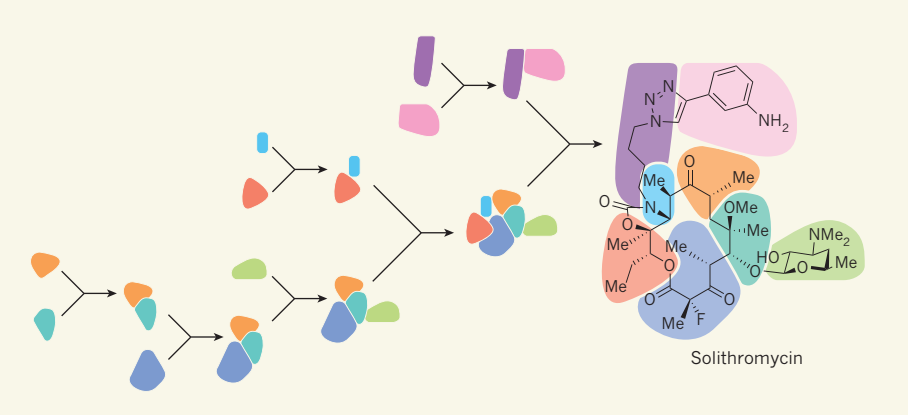


Figure 1 | A modular approach to macrolide synthesis. Seiple *et al.*[1] report a synthetic route that allows the rapid preparation of a diverse array of structural analogues of macrolide antibiotics. The authors prepared eight modular building blocks (coloured shapes), the structures of which can be easily modified. The synthetic route includes seven key steps in which the building blocks are joined together. Chemical groups in the modules can also be readily converted into various other groups after the modules have been incorporated into macrolide intermediates (not shown). This approach allowed the authors to prepare more than 300 macrolides, including the clinical candidate solithromycin.

© 2016 Macmillan Publishers Limited. All rights reserved.

examples highlight the synergy between target-oriented synthesis and methodology developments.

Using such scalable methods allowed the authors to make gram quantities of several macrolides — a large amount for drug-discovery purposes — in remarkably high overall yields. This reaffirms the fact that scalable synthetic methods that can provide ample supplies of structural analogues of molecules are essential for the continued success of drug discovery[4]. Given the sheer number of possible macrolide analogues, some structural variations will remain challenging to produce synthetically, but further advances in synthetic methods will probably bridge the gaps.

Impressively, Seiple and colleagues prepared more than 300 macrolide antibiotics as candidates for pharmaceutical development. The researchers report that most of these compounds inhibit the growth of drug-resistant pathogens, including some extremely challenging strains. Further studies are necessary for assessing properties (such as toxicity and bioavailability) that will affect the antibiotics' clinical viability, but these compounds represent a gateway to commercial pharmaceuticals. One of the compounds is more effective against Gram-negative bacteria than are many other macrolides. Its activity is still modest, but marks a promising start to tackling an important medical need.

No antibiotic is impervious to the power of evolution — bacteria will always mutate to develop resistance, which means that new antibiotics must be pursued continuously and perpetually. To this end, synthetic platforms such as the present one offer rapid access to unparalleled structural variability for antibiotic candidates, while still allowing exquisite control of each variation. This fine-tunability, along with biological knowledge, can enable chemists to keep up with bacterial evolution. In the midst of the antibiotic-resistance crisis, Seiple and colleagues' work sends out a clear message: chemical synthesis is the way forward. ■

Ming Yan *and* **Phil S. Baran** *are in the Department of Chemistry, The Scripps Research Institute, La Jolla, California 92037, USA.*
e-mail: pbaran@scripps.edu

1. Seiple, I. B. *et al. Nature* **533,** 338–345 (2016).
2. Wright, P. M., Seiple, I. B. & Myers, A. G. *Angew. Chem. Int. Edn* **53,** 8840–8869 (2014).
3. Keasling, J. D., Mendoza, A. & Baran, P. S. *Nature* **492,** 188–189 (2012).
4. Parenty, A., Moreau, X., Niel, G. & Campagne, J.-M. *Chem. Rev.* **113,** PR1–PR40 (2013).
5. Kuttruff, C. A., Eastgate, M. D. & Baran, P. S. *Nat. Prod. Rep.* **31,** 419–432 (2014).
6. Seiple, I. B., Mercer, J. A. M., Sussman, R. J., Zhang, Z. & Myers, A. G. *Angew. Chem. Int. Edn* **53,** 4642–4647 (2014).
7. Zhang, Z., Fukuzaki, T. & Myers, A. G. *Angew. Chem. Int. Edn* **55,** 523–527 (2016).

NEUTRINO PHYSICS

A deliberate mix-up in flavour

Neutrinos come in three 'flavours', as do antineutrinos, and they all change flavour as they travel. New measurements of the mixing of different neutrinos may help to explain why our Universe is made of matter and not antimatter.

HELEN O'KEEFFE

For decades, physicists have been intrigued as to why the visible Universe is made of matter, rather than antimatter. Intuitively, studying the smallest of the known fundamental particles, the neutrino, may reveal the answer to this question. As neutrinos travel through space and time, they can oscillate between three different types (flavours). If the way in which antineutrinos change flavour differs from the way neutrinos change flavour, these relatively small differences may have led to a matter-based Universe.

Experiments have been set up to study the amplitude and frequency of these changes to determine the measurable parameters associated with flavour oscillations, such as mass differences (neutrinos have three possible masses, denoted 1, 2 and 3) and mixing angles (how much mixing occurs between neutrinos of mass 1, 2 and 3). Writing in *Physical Review D*, the Daya Bay Collaboration[1] reports that it has used a new detection method to make the most precise measurement yet of θ_{13}, the mixing angle between neutrino mass 1 and neutrino mass 3. This will help in our quest to understand the origin of the dominance of matter in the Universe.

The Daya Bay experiment is located on the southern coast of China, approximately 55 kilometres northeast of Hong Kong. The site consists of eight identical antineutrino detectors, located between 360 m and 1.9 km from the Daya Bay and Ling Ao nuclear power plants. Approximately 3,500 billion billion electron antineutrinos (one flavour of antineutrino) are emitted from the nuclear reactors every second, and a small number of these interact with the 20 tonnes of gadolinium-doped liquid scintillator in each detector; the scintillator emits light when it absorbs an energetic particle. The electron antineutrinos interact with protons in the nuclei of the scintillator molecules to produce positrons (antielectrons) and neutrons in a process known as inverse β-decay. Within a few nanoseconds, the positrons produce light by scintillation, and this is detected by an array of photomultiplier tubes, which amplify the incident light.

When a neutron is captured by a gadolinium nucleus embedded in the scintillator, it forms an excited state, which subsequently decays to a more-stable state by emitting a series of

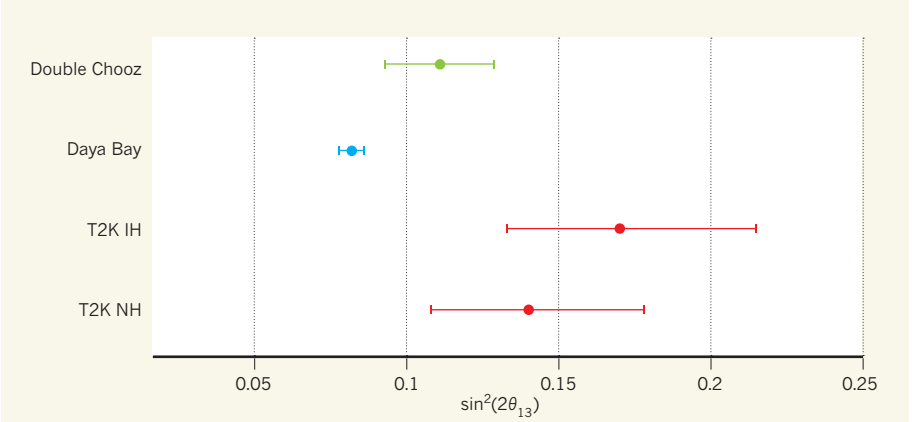


Figure 1 | Neutrino measurements. The mixing angle θ_{13} is a measure of how much neutrinos that have masses 1 and 3 mix. The graph compares measurements of θ_{13} made by the Daya Bay[1], Double Chooz[3] and T2K (ref. 4) experiments; values are plotted as $\sin^2(2\theta_{13})$. Two values are shown for the T2K experiment because the best-fit value of θ_{13} depends on the ordering of the neutrino masses: $m_3 > m_2 > m_1$ is known as the normal hierarchy (NH), whereas $m_2 > m_1 > m_3$ is the inverted hierarchy (IH). The Daya Bay measurement is the most precise so far. A non-zero θ_{13} would open the way to experiments that might explain why the Universe is dominated by matter rather than antimatter.

© 2016 Macmillan Publishers Limited. All rights reserved.

γ-rays with a total energy of 8 megaelectronvolts. These γ-rays scatter and interact with the scintillator to produce another light pulse, typically tens to hundreds of microseconds after the first pulse that was generated by the positrons. Searching for two coincident light pulses that have specific properties can confirm the detection of electron antineutrinos.

Other reactor-based experiments[2,3] use the same gadolinium capture process. But the Daya Bay Collaboration now reports an additional, independent method to identify neutrons produced through inverse β-decay. In the detectors, neutrons are preferentially captured on gadolinium, but capture on a hydrogen atom of the organic liquid scintillator is also possible. Neutron capture on hydrogen produces deuterium and releases a low-energy (2.2-MeV) γ-ray, but this is hard to detect because many similar γ-rays with energies below 4 MeV are emitted from decays of naturally occurring radioisotopes, such as potassium-40 (1.4 MeV) and thallium-208 (2.6 MeV). Such γ-rays are problematic because they can mimic those released by neutron capture on hydrogen, obscuring the antineutrino signals. Radioactivity in the detector must therefore be minimized and controlled by careful purification of the scintillator, selection of low-radioactivity materials and stringent cleanliness. Despite such precautions, radioactivity is never completely eliminated, and this ultimately limits the precision of the experiment.

Detecting inverse β-decay from antineutrino interactions requires the identification of two light signals occurring within 1 to 1,400 microseconds of each other. This timing requirement reduces the impact of coincidental light flashes from background radiation that are not time correlated. The Daya Bay Collaboration was able to estimate the rate of accidental coincidences by using data recorded in its detectors. This quantity, subtracted from the total number of observed coincidences, yields the number of time-correlated coincidences.

Further corrections were required to account for any time-correlated backgrounds, the most significant of which arises from the activation by cosmic rays of carbon-12 in the scintillator to produce lithium-9 and helium-8. Both of these isotopes decay by β-emission to neutron-unstable excited states, producing a signal of an electron followed by neutron capture that is indistinguishable from the signal from inverse β-decay.

When all the backgrounds were appropriately accounted for, the collaboration used the number of genuine time-correlated coincidences to measure the flux of electron antineutrinos. The measured number of electron antineutrinos was found to be less than the collaboration had predicted ***[OK?]***from reactor flux models. The missing electron antineutrinos demonstrate that they oscillate into other flavours that cannot be detected in this experiment.

This analysis from the Daya Bay experiment provides an independent measurement of the number of electron-antineutrino interactions in the detectors and represents an important cross-check of previous measurements of the mixing angle θ_{13}. Moreover, the combination of the collaboration's latest hydrogen-capture results with those from its gadolinium analysis provides the most precise measurement of θ_{13} so far, where $\sin^2(2\theta_{13}) = 0.082 \pm 0.004$, an 8% improvement in precision.

This measurement compares favourably with those from other reactor studies[2,3] and long-baseline experiments[4–7], which provide long distances over which to observe changes in neutrino flavour. Closer examination of the results (Fig. 1) suggests slight discrepancies between current measurements of θ_{13}. Results from Double Chooz[3] and T2K (ref. 4) tend to favour higher values of θ_{13}, although not significantly so.

Measurements of θ_{13} are particularly valuable in our quest to understand the matter-dominated Universe through measurement of another neutrino oscillation parameter, δ_{CP}. Non-zero δ_{CP} would indicate that neutrinos violate charge–parity symmetry (under this symmetry, the laws of physics should be the same when a particle is replaced by its antiparticle and inverted through a mirror); this may explain why the Universe is dominated by matter. Combining results from reactor and long-baseline experiments provides a hint that δ_{CP} is indeed non-zero, but current experiments are not precise enough to reach the statistically significant '5 sigma' discovery threshold. The next generation of experiments[8,9] will seek to make the first measurement of δ_{CP} and thus help us to understand the origin of our matter-dominated Universe. ■

Helen O'Keeffe *is in the Physics Department, Lancaster University, Lancaster LA1 4YB, UK.*
e-mail: h.okeeffe@lancaster.ac.uk

1. An, F. P. *et al.* (Daya Bay Collaboration) *Phys. Rev. D* **93,** 072011 (2016).
2. Ahn, J. K. *et al. Phys. Rev. Lett.* **108,** 191802 (2012).
3. Ishitsuka, M. https://indico.in2p3.fr/event/12279/session/3/contribution/173/material/slides/0.pdf (2016).
4. Abe, K. *et al. Phys. Rev. Lett.* **112,** 061802 (2014).
5. Abe, K. *et al. Phys. Rev. D* **91,** 072010 (2015).
6. Adamson, P. *et al. Phys. Rev. Lett.* **112,** 191801 (2014).
7. Adamson, P. *et al.* Preprint at http://arxiv.org/abs/1601.05022 (2016).
8. Abe, K. *et al. Prog. Theor. Exp. Phys.* **5,** 053C02 (2015).
9. Acciarri, R. *et al.* Preprint at http://arxiv.org/abs/1601.05471 (2016).

REGENERATION

Limb regrowth takes two

Salamanders can regenerate several of their organs, including amputated limbs. Analysis of a Mexican salamander shows that crosstalk between two signalling molecules regulates limb regeneration. SEE LETTER P.407

MIGUEL TORRES

The ability of salamanders to regenerate their limbs following amputation was first described in the eighteenth century by the Italian naturalist Lazzaro Spallanzani[1]. Researchers have been trying to identify the formula for regeneration ever since, in the hope of reawakening this lost ability in mammals. Experiments performed 40 years ago[2] showed that limb regeneration requires the participation of tissues from both sides of the limb — the side closer to the thumb (the anterior side) and the opposite, posterior side. However, the mechanisms involved in this process have remained a mystery. On page 407, Nacu *et al.*[3] show that two diffusible signalling molecules underlie this requirement for anterior and posterior tissue.

During limb regeneration, cells from the limb stump form a mass of progenitor cells called the blastema, which gives rise to the new structure[4]. There are three requirements for this process. First, wounding must occur. This may seem obvious, but wounding triggers several tissue-repair responses that are essential for regeneration. The second requirement is proper innervation, because limbs from which nerves have been removed do not regenerate[5]. Finally, the blastema must contain abutting domains of cells derived from the anterior-most and posterior-most positions in the limb stump[2]. This last observation indicates that crosstalk between cells of differing origins triggers mechanisms that are essential for regeneration.

If wounding on the anterior side of the limb is combined with redirection of nerves from the limb proper into the wounded site, and with transplantation of a patch of posterior limb tissue onto the wound, an extra (accessory) limb can form[6]. By contrast, wounding and nerve redirection without transplantation result in a 'frozen' blastema — regeneration does not occur owing to the lack of

© 2016 Macmillan Publishers Limited. All rights reserved.

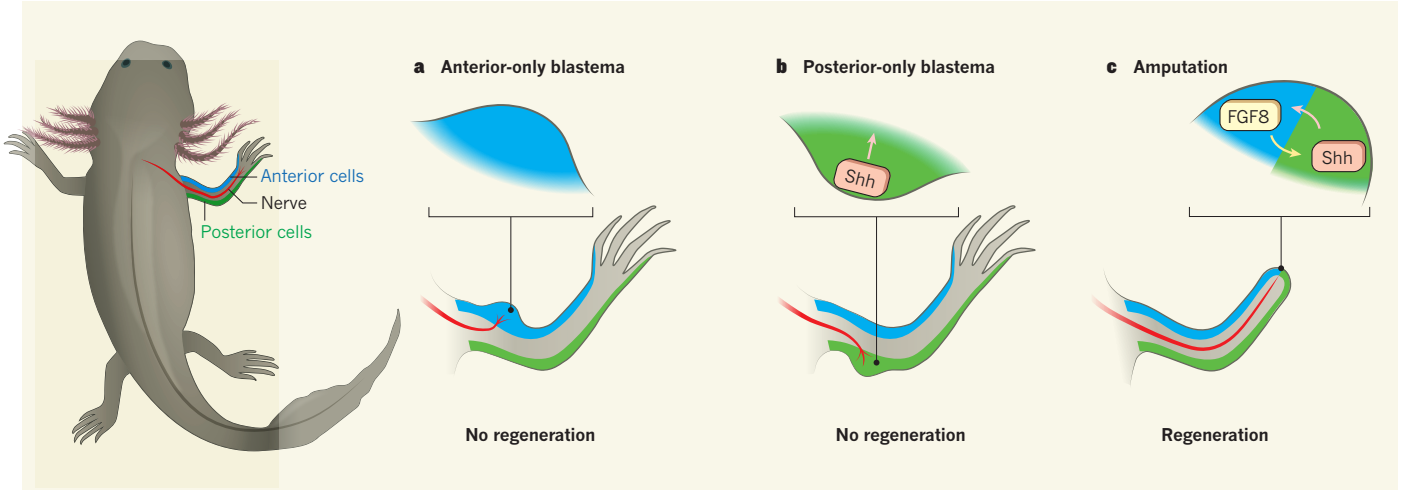


Figure 1 | Complementary signals promote limb regeneration. **a**, In salamanders, if a limb is wounded and nerves are redirected to the wound site, an extra limb can arise from a cell cluster called a blastema. However, the blastema generates a new limb only if it contains cells from both the anterior and posterior sides of the original limb. Nacu *et al.*[3] report that blastemas that contain only anterior tissue cannot form limbs because they do not receive a signalling protein called Shh, which is secreted by the posterior cells. Anterior cells normally produce the signalling protein FGF8 in response to Shh signalling, triggering limb formation. **b**, By contrast, posterior-only blastemas receive Shh, but are unable to activate FGF8 in response. **c**, Following limb amputation, both posterior and anterior cells contribute to the blastema, and regeneration is promoted by cross-induction between Shh and FGF8.

posterior tissue. Using highly regenerative Mexican salamanders called axolotls, Nacu and colleagues took this 'anterior-only' experimental set-up as a starting point to precisely test the molecular pathways involved in antero-posterior crosstalk. The authors found that treating frozen blastemas with a molecule called smoothened agonist (SAG) reawakened the tissue, leading to production of an accessory limb from exclusively anterior tissue.

SAG mimics the activity of a protein called Sonic hedgehog (Shh), which is normally secreted by the posterior-most cells of the blastema. Shh diffuses across the tissue, binding to receptors on the surface of receiving cells to trigger intracellular signalling cascades. Nacu *et al.* found that stimulation of Shh signalling in anterior blastema cells by SAG activated another secreted diffusible protein, fibroblast growth factor 8 (FGF8). Sustained FGF8 activity was necessary for anterior-only limb production (Fig. 1a). Furthermore, activation of FGF8 could induce anterior limb growth even in the absence of Shh signalling.

Like anterior-only blastemas, posterior-only blastemas cannot regenerate, despite containing Shh-expressing cells. Investigating this seeming contradiction, the authors found that posterior tissue could not activate strong and sustained FGF8 signalling in response to Shh (Fig. 1b). Forced FGF8 expression in the posterior tissue led to the production of accessory limbs — but, in this case, only when Shh signalling was also active, because inhibition of Shh blocked regeneration. Finally, Nacu *et al.* confirmed that signalling by Shh and FGF8 was required for physiological limb regeneration after amputation (Fig. 1c).

In summary, the authors have neatly dissected the antero-posterior crosstalk required for regeneration, and have defined the factors that induce blastema development. Anterior cells cannot regenerate alone because they lack Shh signalling from posterior cells. And posterior cells cannot regenerate alone because they are unable to respond to Shh with sustained activation of FGF8. However, why forced FGF8 expression in the absence of Shh signalling cannot promote posterior-only limb growth remains unexplained.

In future analyses, it will be important to determine whether this finding reflects the fact that anterior and posterior tissues have different requirements for regeneration. An alternative possibility is that it reflects experimental design limitations — for example, nonspecific effects of the drug that the researchers used to inhibit Shh in the posterior cells. Similarly, the reason for anterior-specific FGF8 activation by Shh remains unexplained, and should be a focus of future research.

The mechanisms at work during limb regeneration have long been debated, and models fall into two broad classes. Those in the first class posit that every cell contains positional information, and that regeneration involves only local cellular interactions. By contrast, models in the second class invoke the action of diffusible signals produced by specific regions in the blastema called organizers, which direct development of the surrounding tissue.

Nacu and colleagues' work favours organizer models, but does not exclude some degree of positional encoding. Consistent with organizer models, the authors clearly demonstrate that anterior cells do not produce Shh, and that anterior-only accessory limbs do not acquire posterior characteristics (for example, they lack digits that normally arise on the posterior side of the limb). But the authors' data also indicate that blastema cells tend to form sharply defined anterior and posterior domains, suggesting that cells identify themselves according to their position within broad domains — as anterior or posterior, for instance, and, along a perpendicular axis, as dorsal or ventral. In several ways, these observations fit the boundary model[7], which states that organizer regions that produce diffusible signalling molecules are induced at the boundaries between broad limb domains.

Shh and FGF8 have well-defined roles in embryonic limb development in various vertebrate species. A positive feedback loop between the two signalling pathways has been described, and involves relay signalling by Gremlin, a third secreted diffusible protein[8,9]. Nacu *et al.* identified robust Gremlin expression associated with the interaction between anterior and posterior tissue, suggesting that this feedback loop probably also operates during limb regeneration in salamanders.

Given the evolutionary conservation of the Shh and FGF pathways, the current study brings us a step closer to the possibility of using defined factors to promote tissue regeneration in mammals — provided that a blastema can be induced. This advance, together with the same group's recent identification[10] of a diffusible molecule called MLP that is responsible for the cell proliferation leading to blastema formation, represents substantial progress in the quest for a general formula for organ regeneration. ■

Miguel Torres *is in the Cardiovascular Development Program, Centro Nacional de Investigaciones Cardiovasculares, 28029 Madrid, Spain.*
e-mail: mtorres@cnic.es

1. Dinsmore, C. E. *Int. J. Dev. Biol.* **40,** 621–627 (1996).
2. Bryant, S. V. *Nature* **263,** 676–679 (1976).
3. Nacu, E., Gromberg, E., Oliveira, C. R., Drechsel, D. & Tanaka, E. M. *Nature* **533,** 407–410 (2016).

© 2016 Macmillan Publishers Limited. All rights reserved.

4. Nacu, E. & Tanaka, E. M. *Annu. Rev. Cell Dev. Biol.* **27,** 409–440 (2011).
5. Singer, M. *Q. Rev. Biol.* **27,** 169–200 (1952).
6. Lheureux, E. *J. Embryol. Exp. Morphol.* **38,** 151–173 (1977).
7. Meinhardt, H. *J. Embryol. Exp. Morphol.* **76,** 115–137 (1983).
8. Niswander, L., Jeffrey, S., Martin, G. R. & Tickle, C. *Nature* **371,** 609–612 (1994).
9. Zúñiga, A., Haramis, A.-P. G., McMahon, A. P. & Zeller, R. *Nature* **401,** 598–602 (1999).
10. Sugiura, T., Wang, H., Barsacchi, R., Simon, A. & Tanaka, E. M. *Nature* **531,** 237–240 (2016).

This article was published online on 27 April 2016.

ASTROPHYSICS

Illuminating brown dwarfs

Objects known as brown dwarfs are midway between stars and planets in mass. Observations of a hot brown dwarf irradiated by a nearby star will help to fill a gap in our knowledge of the atmospheres of fluid planetary objects. **SEE LETTER P.366**

ADAM P. SHOWMAN

The illumination received from a nearby star has a crucial role in shaping an atmosphere's three-dimensional temperature structure, chemistry, climate and weather. Less obviously, the irradiation of one star by another nearby star — as commonly occurs in tightly orbiting binary star systems — can lead to observable[1] temperature differences between the illuminated star's 'dayside' and its 'nightside'. However, only a few observations have documented the effect of stellar irradiation on the atmospheres of a class of object that is intermediate in mass between stars and planets: brown dwarfs. On page 366 of this issue, Hernández Santisteban *et al.*[2] present intriguing observations to characterize the atmosphere and estimate the day–night temperature difference for a brown dwarf irradiated by a nearby star.

When Sun-like stars reach old age, their outer layers expand, allowing them to engulf nearby, closely orbiting companions. Small stars or brown dwarfs often survive this ordeal, but friction caused by the orbital movement of the companion through the tenuous outer layers of the star provides a drag on the companion, causing it to spiral slowly inward. The outer layers of the bloated star eventually puff off into space, leaving behind a stellar remnant called a white dwarf, which is typically Earth-sized but half as massive as the Sun. By the end of this process, the companion's orbit has often shrunk to the point that the two objects nearly touch.

The system characterized by Hernández Santisteban *et al.*, dubbed J1433, is just such a system, consisting of a white and a brown dwarf. The objects are so close that they orbit each other every 78 minutes. The gravity from the white dwarf distorts the shape of the brown dwarf and leads to a trickle of mass from the companion, which slowly accretes onto the white dwarf.

The closeness of the white and brown dwarfs means that they cannot be resolved individually in images. However, the white dwarf has a temperature that exceeds 13,000 kelvin (more than double the temperature at the surface of the Sun), which causes most of its radiation to escape at short, ultraviolet wavelengths. By contrast, the brown dwarf's surface temperature is about 2,400 K, and most of its radiation escapes in the near-infrared region of the spectrum. Although the white dwarf emits a greater total energy flux, the cooler but larger brown dwarf dominates the system's flux in the near infrared. Radiation in this wavelength range thus allows the brown dwarf's atmosphere to be characterized. Hernández Santisteban *et al.* obtained high-resolution spectra that extended from the ultraviolet to the infrared, allowing the authors to tease apart light from the two objects.

To determine how irradiation affects the brown dwarf, the authors tracked how infrared light from the brown dwarf changes throughout its orbit. The system's orbital plane lies nearly in the line of sight to Earth, implying that the brown dwarf's day and night hemispheres rotate in and out of view throughout the orbit. The researchers' observations show that the average dayside temperature is about 57 K warmer than the average nightside temperature. The hottest dayside region is about 200 K warmer than the coolest nightside region.

These observations are important, given the substantial effort over the past two decades to understand the atmospheres of irradiated exoplanets called hot Jupiters — Jupiter-mass planets that orbit very close to their stars and that are blasted by starlight. Hot Jupiters commonly have daysides many hundreds of kelvins hotter than their nightsides[3,4]. But they are typically about 1,000–10,000 times dimmer than their host stars in the infrared, making their observation extremely difficult. The fact that brown dwarfs in systems such as J1433 are brighter in the infrared than their white-dwarf

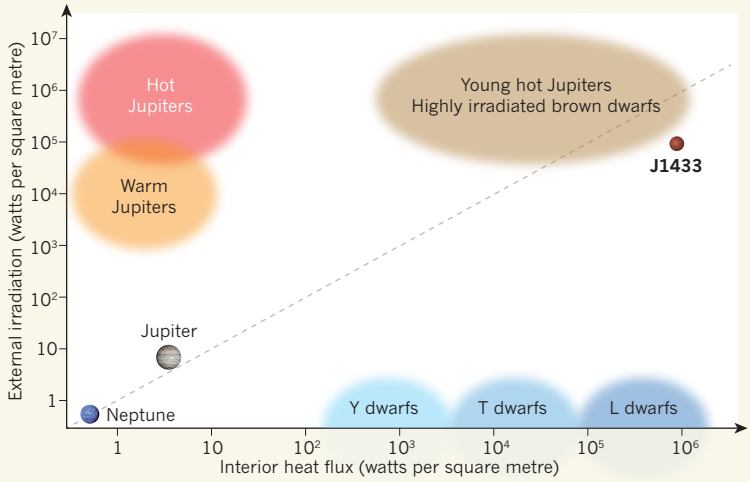


Figure 1 | A wide parameter space for fluid planetary bodies. Fluid planets and brown dwarfs (which are intermediate between planets and stars) are subject to a broad range of internal heat fluxes and external irradiation from nearby stars. Giant planets in our Solar System, such as Neptune and Jupiter, have low internal and external heat fluxes[8]. Warm and hot Jupiters — Jupiter-sized exoplanets — have much higher external irradiation[9], whereas Y-, T- and L-type brown dwarfs have much higher internal fluxes[10]. Hernández Santisteban *et al.*[2] report observations of J1433, a system in which a brown dwarf occupies a tight orbit around a white dwarf star. This brown dwarf has high internal flux and experiences high external irradiation. Observations of other J1433-like brown dwarfs and of young hot Jupiters will therefore provide information that will build a complete picture of atmospheric dynamics in fluid planets and brown dwarfs. Ellipses indicate approximate ranges; the broken grey line indicates equal internal and external fluxes.

© 2016 Macmillan Publishers Limited. All rights reserved.

primaries suggests that the atmospheres of these irradiated objects can be more easily characterized than can those of hot Jupiters, which might allow insight into the workings of the harder-to-observe planets. Several other white dwarf–brown dwarf binaries are known to exist, and may yield constraints on the climate of irradiated fluid objects that are at least as good as those from J1433 (refs 5, 6).

J1433-like systems also allow comparisons with other brown dwarfs. Most known brown dwarfs are isolated and receive no irradiation, so they gradually lose heat from their interiors and cool off over billions of years. This heat is transported through their interiors by convection, which drives an active atmospheric circulation that manifests as patchy, time-variable clouds that cause significant changes in infrared flux over time[7]. Much work is being done to understand this variability and the processes that control the surface patchiness. The extent to which these dynamical processes will be modified by external irradiation is unknown; future observations of J1433 and other irradiated brown dwarfs[5,6] will help to answer this question.

The giant planets in our Solar System (such as Jupiter, Saturn and Neptune) experience internal and external heat fluxes that are weak and comparable to each other[8]. By contrast, hot Jupiters receive external fluxes about a thousand to a million times greater than their expected internal fluxes[9], and thereby show us how atmospheric circulation responds when external forcing dominates. Isolated brown dwarfs represent the opposite extreme, transporting enormous internal fluxes but typically receiving negligible external irradiation. These types of body therefore constrain three corners of a broad parameter space of external irradiation and internal heat flux that spans many orders of magnitude in both parameters (Fig. 1). Until a few years ago, we lacked observational constraints on the atmospheric behaviour of substellar objects at the fourth corner of that parameter space — those subject to enormous external irradiation and internal heat flux that are comparable to within a factor of ten.

J1433 and related brown dwarf–white dwarf binaries fill that gap, and could prove crucial in the quest to understand how atmospheric circulation depends on internal and external forcing. The small day–night temperature difference inferred by Hernández Santisteban *et al.* relative to that of many hot Jupiters[3,4] almost certainly results from the intense heat supplied to the atmosphere from the brown dwarf's interior, but the interaction of the internal and external forcings could have myriad other consequences that remain poorly understood.

The J1433 system is interesting in other ways. Hernández Santisteban *et al.* argue that the brown dwarf began life as a star, but became a brown dwarf after losing mass to the white dwarf — a history that might affect its internal structure and atmospheric circulation. Moreover, because of the fortuitous orbital alignment of J1433 with the line of sight to Earth, the brown and white dwarfs eclipse each other once per orbit, providing an opportunity to characterize the atmospheric composition and thermal structure in the way that is commonly done for hot-Jupiter systems. ■

Adam P. Showman *is in the Department of Planetary Sciences and Lunar and Planetary Laboratory, University of Arizona, Tucson, Arizona 85721, USA.*
e-mail: showman@lpl.arizona.edu

1. Davey, S. & Smith, R. C. *Mon. Not. R. Astron. Soc.* **257,** 476–484 (1992).
2. Hernández Santisteban, J. V. *et al. Nature* **533,** 366–368 (2016).
3. Stevenson, K. B. *et al. Science* **346,** 838–841 (2014).
4. Komacek, T. D. &. Showman, A. P. *Astrophys. J.* **821,** 16 (2016).
5. Steele, P. R. *et al. Mon. Not. R. Astron. Soc.* **429,** 3492–3500 (2013).
6. Casewell, S. L. *et al. Mon. Not. R. Astron. Soc.* **447,** 3218–3226 (2015).
7. Artigau, E., Bouchard, S., Doyon, R. & Lafrenière, D. *Astrophys. J.* **701,** 1534–1539 (2009).
8. Ingersoll, A. P. *Science* **248,** 308–315 (1990).
9. Baraffe, I., Chabrier, G. & Barman, T. *Rep. Prog. Phys.* **73,** 016901 (2010).
10. Marley, M. S. & Robinson, T. D. *Annu. Rev. Astron. Astrophys.* **53,** 279–323 (2015).

STRUCTURAL BIOLOGY

Snapshots of transcription initiation

The enzyme RNA polymerase II, along with several transcription factors, initiates DNA transcription. Analyses reveal the structures involved in this process in human and yeast cells at high-resolution. SEE ARTICLES P.353 & P.359

STEVEN HAHN & STEPHEN BURATOWSKI

The initiation of DNA transcription involves a fascinating interplay between RNA-synthesizing RNA polymerase (Pol) enzymes, transcription factors and DNA. The Pol II complex is of particular interest because it synthesizes all messenger RNA in eukaryotic (nucleus-bearing) cells. The size and flexibility of Pol II complexes present huge challenges for structural biologists, but two studies in this issue, by He *et al.*[1] (page 359) and Plaschka *et al.*[2] (page 353), exploit advances in cryo-electron microscopy to produce near-atomic-resolution snapshots of the Pol II machinery.

The bacterial Pol machinery is a streamlined system that contains only four Pol subunits and a single transcription factor, sigma[3]. Because Pol active sites are highly evolutionarily conserved[4], bacterial Pol has been used to establish a general model of Pol action. This model suggests that Pols and their transcription factors first associate with the promoter region of double-stranded DNA, which lies immediately upstream of the sequences to be transcribed, to form a structure called the closed complex.

Next, around 10–13 base pairs of the promoter unwind, positioning the DNA strand to be transcribed at the Pol active site in an open complex (open and closed refer to the state of the DNA). Pol subunits form channels for incoming nucleotides and the exiting mRNA, and create a deep cleft for the template strand. A mobile clamp domain traps DNA in the active site during the transition from the closed to the open complex. Finally, the structure contorts into an initial transcribing complex, maintaining contacts with promoter DNA while downstream DNA is pulled into the active site as RNA starts to be synthesized.

In comparison with bacteria, the archaeal and eukaryotic transcription machineries are complex, with 12–17 Pol subunits and up to 6 transcription factors. In the Pol II system, the transcription factors TBP, TFIIA, TFIIB, TFIIE, TFIIF and TFIIH are all required and, between them, perform the same functions as bacterial sigma[5,6]. Years of biochemical, molecular and structural studies have probed the roles of each transcription factor to piece together a model of eukaryotic transcription initiation[7].

Many features of this model are brought to life in the current work. Both groups assembled purified transcription factors and Pol II on nucleic-acid scaffolds — on double-stranded promoter DNA for the closed complex, double-stranded DNA containing an unwound 'bubble' for the open complex, and a bubble with a short annealed RNA to resemble the initial transcribing complex. He *et al.* used a complete set of human factors and all three scaffolds, whereas Plaschka *et al.* used all the yeast (*Saccharomyces cerevisiae*) factors except TFIIH to visualize the closed and open complexes. Despite these differences, the positions of transcription factors and the trajectory of the nucleic acids show excellent agreement between the

© 2016 Macmillan Publishers Limited. All rights reserved.

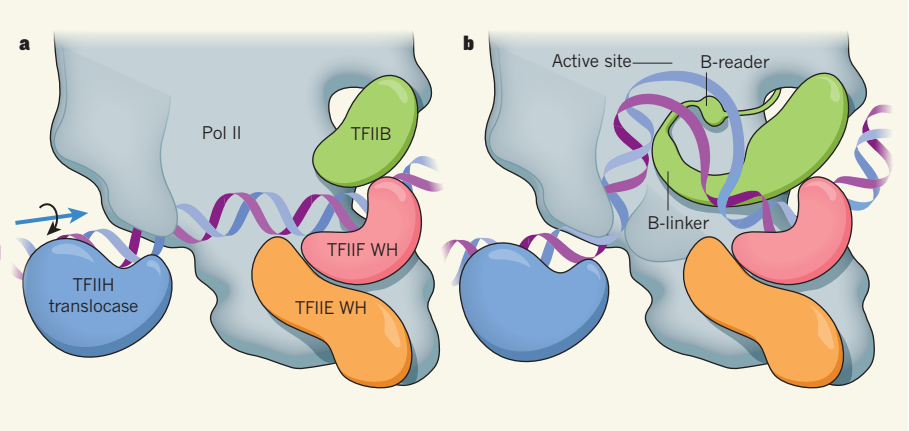


Figure 1 | Preparation for transcription. He *et al.*[1] and Plaschka *et al.*[2] solved structures of the enzyme RNA polymerase II (Pol II) in complex with transcription factors (including TFIIB, TFIIE and TFIIF) and DNA for humans and yeast. They solved closed and open complexes, which form in the lead-up to transcription initiation. **a**, The human closed complex reveals the trajectory of double-stranded DNA across a cleft in Pol II. DNA is enclosed by winged-helix (WH) domains of TFIIE and TFIIF. Arrows show the directions in which DNA is moved by a TFIIH translocase enzyme. A flexible 'B-linker' region of TFIIB (not visible) is disordered in this state. **b**, In the open complex, about 12 bases of DNA have been threaded into the active site of Pol II, leading to the formation of a single-stranded 'bubble' from which the blue strand will be transcribed. Single-stranded DNA is stabilized by interactions with B-linker and B-reader domains of TFIIB. For simplicity, several transcription factors have been removed from this schematic.

studies. Moreover, the structures are consistent with a range of previous analyses[8–12].

TBP recognizes the TATA element — a common DNA sequence in eukaryotic Pol II promoters. A subcomplex of TATA DNA, along with TBP, TFIIA and the cyclin domains of TFIIB, interact with the Pol II wall, positioning downstream promoter DNA over the cleft. A long, flexible region of TFIIB snakes through the Pol II exit channel into the active site, helping to position DNA correctly. In a major advance, the studies reveal how both structured and flexible regions of the transcription factors TFIIE and TFIIF are positioned on either side of the cleft, interacting with each other, Pol II and a TBP–TFIIB–DNA subcomplex to promote and stabilize structural transitions during open-complex formation.

How does the complex transition into an initiation-competent state? Evidence indicates[13,14] that one subunit of TFIIH is a translocase enzyme that drives DNA opening by threading downstream DNA into the Pol II cleft. He and colleagues' structures for the closed and initial transcribing complexes support this model, showing that the translocase maintains contact with DNA in both states, and that an extra 12 bases are threaded into the cleft by the time that the latter has formed.

The two groups present slightly different models for the transition from the closed to the open complex. He *et al.* propose that, during the transition, closure of the clamp — driven by translocase-generated torsional strain — traps single-stranded DNA. Clamp-associated TFIIE also shifts, and a flexible linker region in TFIIB is fixed to stabilize single-stranded DNA (Fig. 1). By contrast, Plaschka *et al.* propose that, in yeast, structural changes mediated by TFIIE–TFIIB contacts clear the cleft of flexible TFIIB segments, allowing entry of the template strand, followed by closure of the clamp. In this model, TFIIE, and probably TFIIB, stabilize single-stranded DNA in the open complex.

Discrepancies between the models could reflect species differences in transcription factors or promoter sequences, differences in the size of the DNA bubble, or the absence of TFIIH in the yeast system. Both models involve small changes in flexible segments of TFIIB, TFIIE and TFIIF, so, despite the remarkable resolution achieved, the structures might still have insufficient resolution to identify subtle differences. Moreover, these static structures capture only single states of a dynamic, multi-step mechanism. Perhaps the torsional strain that drives these transitions affects the structural state of the complexes — an area for future study.

The current research highlights similarities and differences in initiation mechanisms between all multi-subunit Pols. For instance, bacterial-DNA unwinding is initiated when a wedge structure in sigma flips out a base, leading to the opening of the DNA bubble[15,16]. TFIIB, which has a completely different amino-acid sequence from sigma, interacts with the same segment of promoter DNA. Both factors stabilize the open state and must be ejected from the active site before transcription progresses into elongation mode. Furthermore, eukaryotic Pols I, II and III all use TBP, a TFIIB-like factor and subunits related to TFIIE and TFIIF (ref. 17), but Pol II is the only one to require a translocase. The reason for this is unknown, but probably involves greater stability of the Pol I and III open complexes, analogous to the stable bacterial open complex.

These fantastic new structures provide a framework for interpreting many past results, and are a harbinger of future investigations. For instance, a recent structure[18] of a TBP-containing complex called TFIID, which helps to activate transcription, revealed that some TFIID subunits interact with promoter DNA. This large structure can be docked into the current studies' closed-complex structures — suggesting unexpected interactions between the basal transcription factors and TFIID subunits that may be important in transcriptional regulation.

Eventually, it should be possible to directly visualize how multi-subunit coactivator complexes, such as TFIID, or another coactivator, Mediator, interact with the transcription machinery. Mediator is missing from the authors' structures, but is required for nearly all Pol II transcription[19]. The carboxy-terminal domain of Pol II's largest subunit, a key regulatory target[20], is also missing. Rapidly advancing technologies for structural biology hold great promise for many more interesting insights into the regulation of gene expression. ■

Steven Hahn *is at the Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA.* **Stephen Buratowski** *is in the Department of Biochemical Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02115, USA.*
e-mails: shahn@fredhutch.org; steve_buratowski@hms.harvard.edu

1. He, Y. *et al. Nature* **533,** 359–365 (2016).
2. Plaschka, C. *et al. Nature* **533,** 353–358 (2016).
3. Saecker, R. M., Record, M. T. & deHaseth, P. L. *J. Mol. Biol.* **412,** 754–771 (2011).
4. Cramer, P. *et al. Annu. Rev. Biophy.* **37,** 337–352 (2008).
5. Werner, F. & Grohmann, D. *Nature Rev. Microbiol.* **9,** 85–98 (2011).
6. Thomas, M. C. & Chiang, C.-M. *Crit. Rev. Biochem. Mol. Biol.* **41,** 105–178 (2006).
7. Sainsbury, S., Bernecky, C. & Cramer, P. *Nature Rev. Mol. Cell Biol.* **16,** 129–143 (2015).
8. Sainsbury, S., Niesser, J. & Cramer, P. *Nature* **493,** 437–440 (2013).
9. He, Y., Fang, J., Taatjes, D. J. & Nogales, E. *Nature* **495,** 481–486 (2013).
10. Murakami, K. *et al. Proc. Natl Acad. Sci. USA* **112,** 13543–13548 (2015).
11. Grünberg, S., Warfield, L. & Hahn, S. *Nature Struct. Mol. Biol.* **19,** 788–796 (2012).
12. Kim, T.-K., Ebright, R. H. & Reinberg, D. *Science* **288,** 1418–1421 (2000).
13. Fishburn, J., Tomko, E., Galburt, E. & Hahn, S. *Proc. Natl Acad. Sci. USA* **112,** 3961–3966 (2015).
14. Fazal, F. M., Meng, C. A., Murakami, K., Kornberg, R. D. & Block, S. M. *Nature* **525,** 274–277 (2015).
15. Zhang, Y. *et al. Science* **338,** 1076–1080 (2012).
16. Bae, B., Feklistov, A., Lass-Napiorkowska, A., Landick, R. & Darst, S. A. *eLife* **4,** e08504 (2015).
17. Vannini, A. & Cramer, P. *Mol. Cell* **45,** 439–446 (2012).
18. Louder, R. K. *et al. Nature* **531,** 604–609 (2016).
19. Poss, Z. C., Ebmeier, C. C. & Taatjes, D. J. *Crit. Rev. Biochem. Mol. Biol.* **48,** 575–608 (2013).
20. Buratowski, S. *Mol. Cell* **36,** 541–546 (2009).

This article was published online on 11 May 2016.

© 2016 Macmillan Publishers Limited. All rights reserved.

doi:10.1038/nature17987

Reproducible pharmacogenomic profiling of cancer cell line panels

Peter M. Haverty[1]*, Eva Lin[2]*, Jenille Tan[2], Yihong Yu[2], Billy Lam[2], Steve Lianoglou[1], Richard M. Neve[2], Scott Martin[2], Jeff Settleman[2], Robert L. Yauch[2] & Richard Bourgon[1]

The use of large-scale genomic and drug response screening of cancer cell lines depends crucially on the reproducibility of results. Here we consider two previously published screens, plus a later critique of these studies. Using independent data, we show that consistency is achievable, and provide a systematic description of the best laboratory and analysis practices for future studies.

Human cancer cell lines are vital models for studying anti-cancer therapeutics and determinants of drug sensitivity. With the rapidly growing recognition of intra-tumour and inter-patient heterogeneity, the optimal use of these models demands a comprehensive characterization of their molecular features and a sufficient number of models to capture the genomic diversity observed across human cancers. Two recent projects, the Genomics of Drug Sensitivity in Cancer (GDSC) and the Cancer Cell Line Encyclopedia (CCLE), collectively characterized more than 1,000 cell lines and nearly 150 anti-cancer agents[1,2], providing a potentially valuable resource for the research community. However, a subsequent meta-analysis published in this journal asserted that drug response results failed to agree[3]. As a consequence, Haibe-Kains *et al.*[3] expressed serious concerns about the validity and use of these resources, thus raising questions about screening projects of this type more generally. To address these concerns and to determine the causes for the controversy, we independently characterized the response of 410 cancer cell lines to 16 agents tested by GDSC and CCLE. We also evaluated three specific aspects of the screening protocols that are relevant to measured drug response: readout of cell viability (metabolic versus DNA content), seeding density strategy, and cell culture media conditions. For screens of this type to be useful, they must accurately identify drug-sensitive and drug-resistant cell lines, as well as biomarkers of drug response that generalize to other contexts and data sets. Using more robust viability statistics and concordance metrics that are aligned with these two key screening objectives, we showed strong agreement between our own results and those generated by CCLE, thereby demonstrating that reliability is achievable. Concordance with GDSC was weaker, although still appreciable. Our data further demonstrated that, for several drugs, differences in media conditions and seeding density contributed to residual differences between studies. Unexpectedly, we found that metabolic and DNA content readouts produced very similar estimates of drug sensitivity, but that the DNA content readout was markedly noisier. Taken together, our results revealed the technical and analytical parameters that are crucial for achieving replicable results. These findings should be of value in ensuring the robustness of data generated by similar efforts, and in maximizing the utility of cancer cell line panels for the study of anti-cancer agents and biomarkers of drug sensitivity.

Results

The Genentech Cell Line Screening Initiative (gCSI) was undertaken independently of CCLE and GDSC. Nonetheless, gCSI had profiled 410 cell lines and 16 drugs that were also evaluated by the other projects. Figure 1 and Extended Data Fig. 1 compare the mean of fitted viability values (equivalent to the familiar area under the curve, Methods) as well as 50% inhibitory concentration (IC_{50}, Methods) statistics produced by gCSI with those from CCLE or GDSC. As expected, chemotherapeutics typically inhibited growth of nearly all cell lines to some extent, while targeted agents such as the selective kinase inhibitors erlotinib or lapatinib only affected a small number of cell lines that are presumably dependent on specific pathways.

Identifying sensitive or resistant cell lines

Figure 1 and Extended Data Fig. 1 provide a useful overview of the data as well as continuity with the presentation of Haibe-Kains *et al.* (figure 2 in ref. 3). However, this representation of drug response has two crucial shortcomings that, if unaddressed, confound cross-study comparisons. First, truncation to tested doses (for IC_{50}) or to the 0–100% range (as done by GDSC for mean viability) obscures important aspects of the data distribution and yields large numbers of ties, which are not appropriate for many standard statistical approaches. Second, although IC_{50} is a familiar and interpretable statistic and is often appropriate for smaller screens, it is not well-suited to high-throughput screening. For the large number of cell lines that never achieve 50% inhibition, it forces either truncation or extrapolation, both of which are potentially problematic. Furthermore, it can produce high-variance estimates (see Extended Data Fig. 2 for illustration). The mean viability statistic avoids both of these problems, and for this reason, we use it exclusively in all analyses that follow.

Because high-confidence identification of groups of sensitive and resistant cell lines is a primary objective of profiling, we first compared gCSI and CCLE results on this basis. For targeted agents expected to inhibit only a few lines (Fig. 2a), we fit two-component mixture distributions to assign 'sensitive', 'no-call' or 'resistant' labels (Extended Data Fig. 2, Methods). For broadly active compounds (Fig. 2b), it was more appropriate to divide lines into the three response categories in equal proportion. It is common to omit unclassified lines or those exhibiting intermediate response; in cases where both data sets confidently assigned a sensitive or resistant label, the strong majority of calls were concordant (84–96%, Table 1). Furthermore, cell lines with molecular features previously associated with drug response (for example, *ERBB2* overexpression for lapatinib) were almost entirely absent from regions where one or both data sets assigned a resistant label (Fig. 2). Thus, independently generated

[1]Department of Bioinformatics and Computational Biology, Genentech Inc., 1 DNA Way, South San Francisco, California 94080, USA. [2]Department of Discovery Oncology, Genentech Inc., 1 DNA Way, South San Francisco, California 94080, USA.
*These authors contributed equally to this work.

© 2016 Macmillan Publishers Limited. All rights reserved

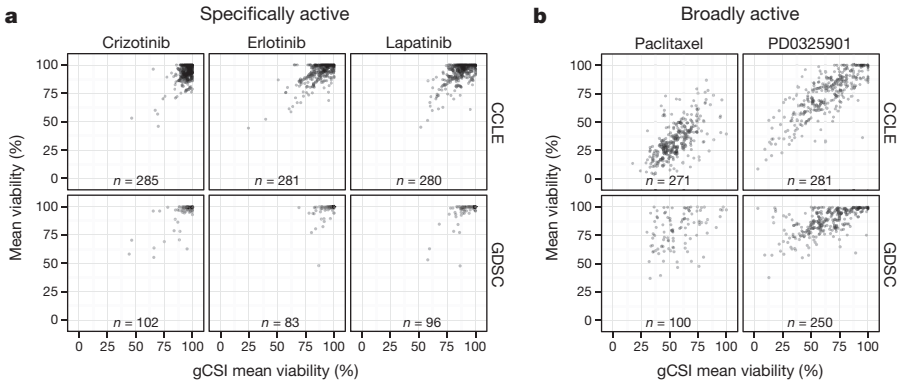


Figure 1 | gCSI, CCLE and GDSC drug response statistics reveal different patterns of response for specifically versus broadly active drugs. We include all lines in the two-way gCSI/CCLE overlap in the top rows, and all lines in the two-way gCSI/GDSC overlap in the bottom rows (see Extended Fig. Data 3 for restriction to only those lines included in the three-way overlap). In each panel, *n* refers to the number of lines in common and plotted for each pairwise comparison. Mean viability (equivalent to area under the log-dose/response curve) across tested drug concentrations for cell lines as assayed by each pair of projects is plotted for specifically active drugs (**a**) and broadly active drugs (**b**). 0% corresponds to complete growth inhibition at all doses; 100% corresponds to viability equal, on average, to untreated control wells. Values are capped at a maximum value of 100% for consistency with GDSC published data. gCSI values are derived from three biological replicates (Methods).

drug response data can reveal very similar sets of high-confidence sensitive and resistant lines.

When gCSI and GDSC results were compared on a similar basis, concordance for PD0325901 and paclitaxel was good, although not as strong as that observed relative to CCLE (Table 1). This was due in part to the lower maximal doses used by the GDSC, and their truncation of mean viability at 100%. For two out of three targeted agents, on the other hand, gCSI was in slightly higher agreement with GDSC than with CCLE (Table 1). However, concordance here was driven by the predominance of resistant lines, since GDSC seems to have selected very few cell lines sensitive to erlotinib and lapatinib (Extended Data Fig. 3). For the 10 drugs in common between only gCSI and GDSC, concordance between these groups was also generally lower than for the five drugs common to all three groups (Extended Data Fig. 4).

Genomic biomarkers for drug response

We next turned to a second concordance metric aligned with a key screening objective: identification of predictive biomarkers of drug response. In their original publications, both CCLE and GDSC used a well-established penalized regression strategy, the 'elastic net'[4], to identify predictive biomarkers[1,2]. The mathematics behind this method guarantees that only a handful of the tens or hundreds of thousands of candidate molecular features are selected. For many applications, such parsimony is highly desirable. When contrasting two sets of results, however, it can make assessment of underlying similarity difficult, because some features that are in fact well-correlated with response may be omitted if others that provide equivalent information are selected instead. A similar phenomenon has been described when constructing gene expression signatures to assign cancer subtype[5]. Despite this caveat, a direct comparison of elastic

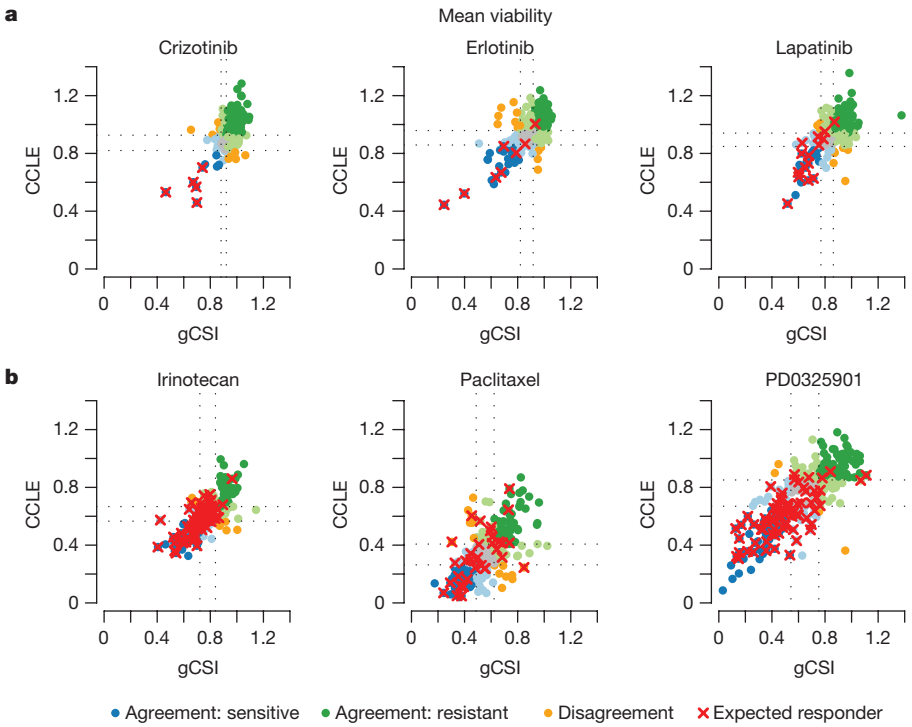


Figure 2 | Highly concordant sensitive and resistant labels assigned via mean viability. a, For targeted agents, dotted lines indicate boundaries between sensitive, no-call and resistant lines (Methods). Expected responder cell lines (those with molecular features known to be associated with drug response) are indicated in red for each drug (Methods). **b**, For broadly active compounds, lines were divided into three equally sized groups on the basis of their ordered mean viability: sensitive, no-call and resistant.

© 2016 Macmillan Publishers Limited. All rights reserved

Table 1 | Classification concordance

		CCLE						GDSC					
			Agree						Agree				
Type	Drug	Called (%)	Sensitive	Resistant	Disagree	Agree (%)	*P* value	Called (%)	Sensitive	Resistant	Disagree	Agree (%)	*P* value
Specific	Crizotinib	60	9	146	17	90	2.0×10^{-6}	85	9	68	10	89	5.8×10^{-6}
	Erlotinib	57	35	100	25	84	5.8×10^{-16}	66	3	48	4	93	2.9×10^{-3}
	Lapatinib	48	24	100	10	93	3.1×10^{-17}	59	2	54	1	98	1.9×10^{-3}
Broad	Irinotecan	53	41	38	9	90	1.0×10^{-14}	–	–	–	–	–	–
	Paclitaxel	47	57	55	16	88	1.5×10^{-18}	60	23	19	18	70	3.9×10^{-3}
	PD0325901	50	72	64	5	96	9.2×10^{-34}	46	55	45	16	86	8.7×10^{-16}

Lines called sensitive or resistant by both gCSI and CCLE (or separately, by gCSI and GDSC) were assessed for concordant classification. The percentage of such doubly called lines is given in the 'Called' column. The counts of cell lines called sensitive or resistant by both groups, or with discordant classifications, are listed along with the percentage of doubly called cell lines with concordant classifications. *P* values correspond to a Fisher's exact test for the sensitive/resistant two-way table. For targeted agents, total concordance was slightly higher for GDSC than CCLE, although very few sensitive lines were called among the gCSI/GDSC overlap. For broadly active agents, total concordance was significantly higher for CCLE.

net results obtained from gCSI and CCLE data revealed substantial similarity. Specifically, for five of the six drugs common to the studies, known correlates of response were recovered from gCSI and CCLE response data when fit against either their own genomic data or that of the other study (Extended Data Fig. 5a). gCSI was also able to recover a known correlate of crizotinib response when fit against its own genomic data (Extended Data Fig. 5a).

To perform a more sensitive comparison, we next applied the elastic net to one study's drug response data to select biomarkers, and then separately, applied univariate regression to score all candidate biomarkers relative to the other study's response data (features that scored strongly had predictive value for response). Figure 3 shows that although a small number of features selected by one study appeared uninteresting in the other, the vast majority of candidate biomarkers selected by either gCSI or CCLE also produced extreme association statistics when compared to the other study's drug response data. For example, *ERBB2* and *CDH3* were among the genes selected by the elastic net when applied to gCSI response data for lapatinib; univariate analysis of CCLE response data produced false discovery rate (FDR)-corrected *P* values of 1.6×10^{-12} and 6.6×10^{-7} for these genes, placing them on the far left of the CCLE distribution. Use of discrete sensitive/resistant labels rather than continuous drug response produced qualitatively similar results (Extended Data Fig. 6).

Owing to the truncation of published GDSC mean viability data, gCSI could only be compared to GDSC on the basis of biomarker discovery using discrete response. Although gCSI and CCLE discretized response data permitted identification of known correlates of response for both specific and broadly active drugs when assessed on lines in common between gCSI and CCLE (Extended Data Fig. 6a), only the latter were recovered when the same analysis was run on lines in common between gCSI and GDSC (Extended Data Fig. 7a). Furthermore, for drugs with specific effect, the algorithm yielded no biomarkers at all for GDSC drug response data when run on lines in common between gCSI and GDSC. For ten additional drugs in common between only gCSI and GDSC, the elastic net performed better—both gCSI and GDSC drug response data were capable of producing elastic net hits that scored well in the other study for many drugs in this set (Extended Data Fig. 4a, b).

Methodology contributes to discrepancy

After demonstrating that independent data sets were capable of revealing common responsive and resistant lines and common sets of candidate biomarkers, we next characterized potential sources of systematic variation between the projects. Comparison of experimental methods highlighted three key differences: assay type, cell seeding density, and growth media. The specific choices of each project are described in Extended Data Table 1a.

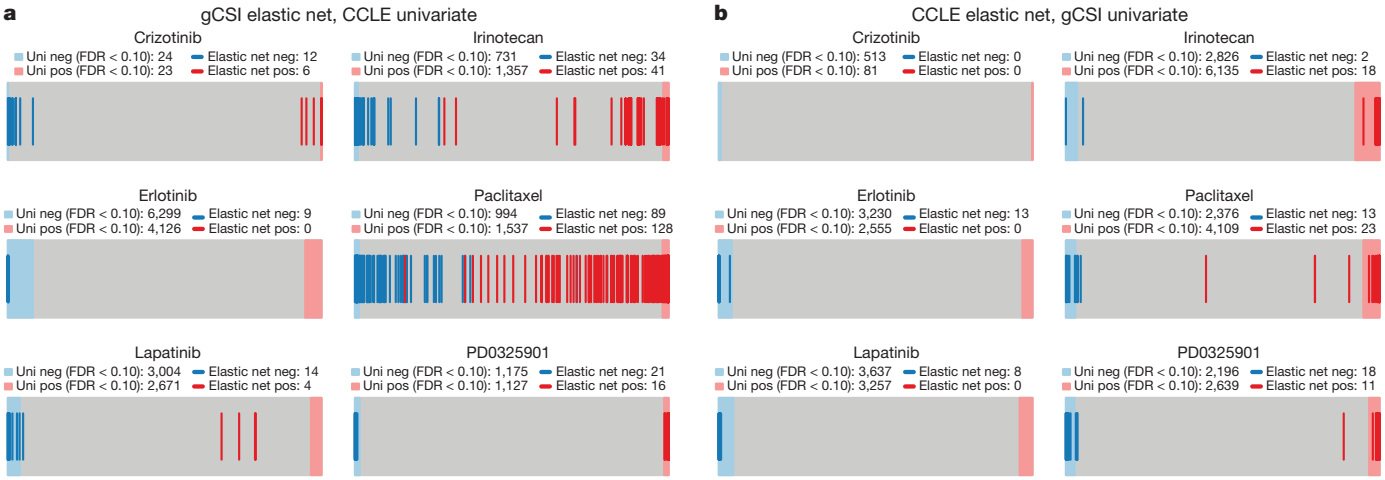


Figure 3 | Consistent biomarker scoring for gCSI and CCLE drug response data. The elastic net algorithm was applied to identify short lists of candidate biomarkers for drug response as assessed by either gCSI or CCLE mean viability values, using gCSI genomic data. (Equivalent results based on CCLE genomic data are shown in Extended Data Fig. 5b–d.) Supplementary Table 4 describes the number of cell lines available for each comparison. Extended Data Fig. 5a describes the recovery of biomarkers expected to associate with response for each of the six drugs tested by both gCSI and CCLE. **a**, Genomic features were ranked by direction and strength of their association with CCLE mean viability (from strongest negative to strongest positive). Shaded regions indicate features with significant univariate association (FDR-adjusted $P<0.10$). Blue indicates negative association with response and red indicates a positive association. The univariate ranks of features selected by the elastic net using gCSI drug response are indicated with vertical bars, also coloured by direction of association. In general, the gCSI-selected features are strongly enriched at the extreme ends of the distributions, indicating that they are highly relevant to both drug response data sets. **b**, As for panel **a**, but applying the elastic net to CCLE drug response data and comparing results to univariate statistics obtained from gCSI drug response.

© 2016 Macmillan Publishers Limited. All rights reserved

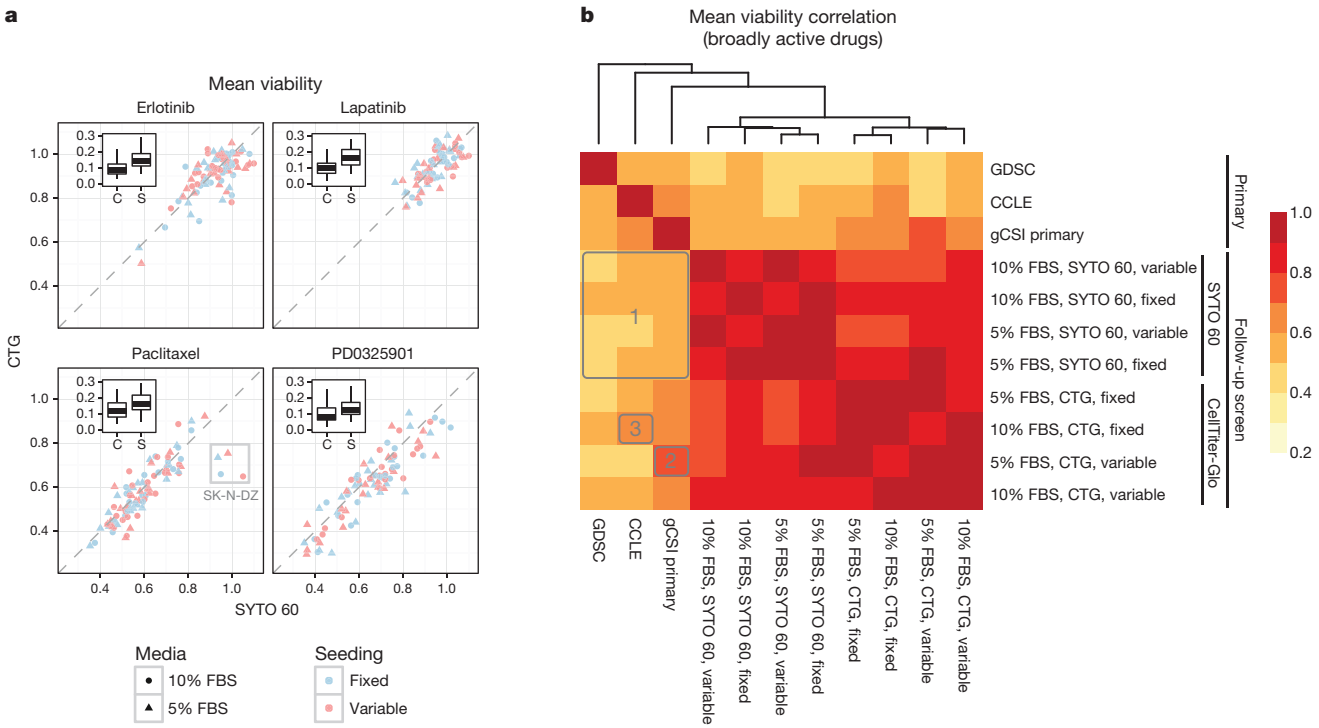


Figure 4 | Follow-up screen shows SYTO 60 assay closely agrees with CTG but is significantly less precise. a, Comparison of mean viability between STYO 60 and CTG assays for 24 lines and four combinations of growth media and seeding density. Results were largely concordant, with SK-N-DZ treated with paclitaxel as the only notable exception. Symbol colour indicates seeding strategy; symbol shape indicates media condition. Inset plots indicate the widths of inter-replicate 95% confidence intervals on mean viability for CTG (C) and SYTO 60 (S). See Extended Data Fig. 8 for full details. **b**, Pairwise Pearson correlation of mean viability across 24 cell lines, combining data from the two broadly active drugs, for which use of correlation coefficient is appropriate. Follow-up SYTO 60 results were more consistent with CTG values from the primary gCSI and CCLE screens than with the GDSC SYTO 60 values (rectangle '1'). All eight follow-up screen results were best correlated with the gCSI primary screen, with the strongest correlation (rectangle '2', $r = 0.74$) being the same conditions used in the primary gCSI screen. Similarly, follow-up results that best correlated with CCLE data were those that matched CCLE screening conditions (rectangle '3', $r = 0.63$). CTG and SYTO 60 values are derived from three biological replicates (Methods).

We were particularly interested in the potential impact of the choice of assay for viability readout. Cytotoxic effects can be reasonably assessed by either metabolic readout or live cell count, whereas cytostatic effects may produce distinct results in the two viability readouts[6]. For lapatinib and paclitaxel, for example, Haibe-Kains *et al.*[3] observed a better correlation between CCLE and a GlaxoSmithKline (GSK) study[7] that also used CellTiter-Glo (CTG) than between GDSC and GSK. On this basis they asserted that inconsistencies between CCLE and GDSC were partly due to the use of different assays[3].

To evaluate the relevance of methodology differences, we undertook a follow-up screen that considered each of the variables in a full three-way factorial design. Twenty-four lines and four drugs common to all three studies were re-assessed with CTG versus the SYTO 60 fluorescent stain, fixed versus variable seeding, and 5% versus 10% FBS (Methods). PD0325901 showed a slight but systematic reduction in viability when assessed by CTG (mean reduction of 3.6 percentage points relative to SYTO 60, paired *t*-test excluding SK-N-DZ cells, $P = 1.3 \times 10^{-5}$, Fig. 4a). Notably, for the other three drugs, a comparison of CTG and SYTO 60 mean viability estimates showed no evidence of systematic bias (erlotinib, $P = 0.77$; lapatinib, $P = 0.23$; and paclitaxel, $P = 0.27$). SK-N-DZ cells treated with paclitaxel were the only exception; however, since this line is only loosely adherent, it may have been subject to cell loss during a SYTO 60 washing step. Further, our SYTO 60 results for broadly active drugs (including PD0325901) were more consistent with CTG values from the primary gCSI ($r = 0.58$) and CCLE ($r = 0.47$) screens than with the GDSC SYTO 60 values ($r = 0.42$, Fig. 4b). Taken together, these data indicate that, at least for the drugs considered in the follow-up screen, biological bias arising from choice of viability readout is not a plausible explanation for the differences between CCLE and GDSC. Crucially, results from the SYTO 60 assay were significantly more variable than those from CTG. Confidence intervals for mean viability (Methods) were significantly wider for SYTO 60 than for CTG—for most cell lines, and for all drugs and across all media and seeding density conditions (Fig. 4a and Extended Data Fig. 8c). Combining these observations, we see that the SYTO 60 assay typically captured the same information as CTG but it did so with far less precision.

Unlike viability readout, other methodology choices did impact results more broadly, although typically with modest effect size. Extended Data Figure 8a shows that for some, although not all, drug and cell line combinations, increased FBS had a consistent though weak effect on mean viability. For PD0325901 in particular, increasing FBS systematically increased mean viability (an average increase of 4 percentage points, paired *t*-test, $n = 46$, $P = 2.3 \times 10^{-5}$, Extended Data Fig. 8b). For other drugs, however, no consistent shift of a practically relevant magnitude was observed (Extended Data Fig. 8b).

Seeding density also produced systematics effects. When slow growing cell lines were seeded at higher density—allowing control wells to become confluent over the course of the assay, and thus potentially constraining growth or saturating CTG signal—mean viability tended to increase and apparent drug sensitivity was reduced. Conversely, fast growing cell lines tended to show lower mean viability and appeared to be more sensitive when seeded at lower density (Extended Data Fig. 9a). Intriguingly, a similar trend was seen when comparing gCSI (variable seeding) and CCLE (fixed seeding) mean viability data (Extended Data Fig. 9b), suggesting that a small but statistically significant portion of the difference between the two result sets can be explained by how those projects managed differential growth rates.

© 2016 Macmillan Publishers Limited. All rights reserved

Discussion

In their critique of the CCLE and GDSC projects, Haibe-Kains *et al.*[3] wrote, "Because there is no clear concordance, predictive models of response developed using data from one study are almost guaranteed to fail when validated on data from another study, and there is no way with available data to determine which study is more accurate." The authors of the original studies recently disputed the choice of statistical methods used in the critique, and in suggesting alternatives, they highlighted the importance of biologically grounded analysis methods[8]. We strongly endorse this point of view. Further, we have presented an important, independent data resource and shown that when robust drug response statistics are used, and when concordance metrics are aligned with the stated purpose of high-throughput cell line drug response profiling, consistency can be achieved: consistent identification of cell lines that are sensitive or resistant to drug, as well as consistent identification of candidate biomarkers that predict response.

Furthermore, the independently generated gCSI data address the question of which of the two original studies is more accurate: gCSI data agreed less well with GDSC than with CCLE. The GDSC truncation of viability statistics as well as lower dose ranges and limited sets of sensitive cell lines for some drugs undoubtedly contributed to this difference. We have also shown that, in our hands, SYTO 60 produced significantly noisier drug response estimates than CTG, even while capturing essentially the same biology. Therefore, the GDSC choice of SYTO 60 probably further reduced concordance with gCSI (and CCLE).

Finally, we have shown that other aspects of experimental methodology systematically affect measured cell viability. In particular, allowing cells to become over-confluent, by either over-seeding or using richer media, can lead to underestimation of inhibition. While these phenomena explain some fraction of the discrepancy between projects, other sources of both stochastic and unexplained systematic error undoubtedly remain. Our data demonstrate, however, that when appropriate laboratory and analysis protocols are applied, these are small enough to permit the generation of biomarker results that replicate in other settings, and thus, a realization of the potential of large cell line panels for characterization of heterogeneous response to anti-cancer agents and therapeutics.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 24 October 2014; accepted 7 April 2016.

1. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483,** 603–607 (2012).
2. Garnett, M. J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483,** 570–575 (2012).
3. Haibe-Kains, B. *et al.* Inconsistency in large pharmacogenomic studies. *Nature* **504,** 389–393 (2013).
4. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33,** 1–22 (2010).
5. Ein-Dor, L., Kela, I., Getz, G., Givol, D. & Domany, E. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* **21,** 171–178 (2005).
6. Chan, G. K. Y., Kleinheinz, T. L., Peterson, D. & Moffat, J. G. A simple high-content cell cycle assay reveals frequent discrepancies between cell number and ATP and MTS proliferation assays. *PLoS ONE* **8,** e63583 (2013).
7. Greshock, J. *et al.* Molecular target class is predictive of *in vitro* response profile. *Cancer Res.* **70,** 3677–3686 (2010).
8. Cancer Cell Line Encyclopedia Consortium & Genomics of Drug Sensitivity in Cancer Consortium. Pharmacogenomic agreement between two cancer cell line data sets. *Nature* **528,** 84–87 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank G. Manning, T. Sandmann, B. Forrest and D. Stokoe for their valuable contributions to improving the manuscript. We also thank L. Shi and J. Wu for work on early phases of gCSI screening; R. Rodriguez, G. Yuen and D. Hascall for help preparing drug plates; Y. Jiang for assistance with drug plate quality control; and S. Selvaraj and M. Yu for banking and quality control of cell lines used in this study. This manuscript contains an analysis of data released by the Broad Institute (CCLE) and by the GDSC members. Those who carried out the original analysis and collection of these data bear no responsibility for the further analysis or interpretation of it. The relevant subset of the CCLE data are included in the supplementary software package with the explicit written consent of the CCLE group. The GDSC data has been included following the GDSC's instructions for attribution and under their specified open-source license, Creative Commons 3.0.

Author Contributions P.M.H. and R.B. drafted the manuscript, prepared figures and tables and designed the factorial screen. P.M.H. and R.B. performed computational analysis and interpreted results. S.L. contributed to the development of elastic net biomarker identification software. J.S. and R.L.Y. contributed to the experimental design, data analysis, and manuscript preparation. S.M. contributed to manuscript preparation and data interpretation. E.L., J.T., Y.Y. and B.L. performed primary cell-based screening experiments. E.L. and R.M.N. designed and managed the cell-based screening experiments.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.B. (bourgon.richard@gene.com).

© 2016 Macmillan Publishers Limited. All rights reserved

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Cell lines. Genentech maintains a collection of cancer cell lines obtained from a variety of academic and commercial sources (for example, ATCC, DSMZ)[9]. Cell line identity was verified by high-throughput single nucleotide polymorphism (SNP) genotyping using Illumina Golden gate multiplexed assays. SNP profiles were compared to SNP calls from internal and external databases to determine or confirm ancestry. In cases where data were unavailable or cell line ancestry was questionable, DNA or cell lines were re-purchased to enable confirmation of cell line ancestry. SNP genotypes were assessed each time new stocks were expanded from cryopreservation. SNP genotypes were also matched to our SNP microarrays (see below), which were subsequently matched to our RNA-Seq and exome sequencing data to confirm the matching of samples used for genomic and cell-based assays.

Screening drugs management and quality control. Drugs were obtained from in-house synthesis or purchased from commercial vendors. A fully automated transfer system by Nova Technology (Innovate Engineering) was used to transfer material from a dry library, solubilize with DMSO, and log the solutions into our compound management system. A high throughput liquid chromatography mass spectrometry/ultraviolet absorbance/charged aerosol detector/chemiluminescent nitrogen detector (LCMS/UV/CAD/CLND) system was used to verify identity, purity, and concentration of drugs used in the gCSI screens. The LCMS/UV/CAD/CLND system consisted of an LCMS/UV system (Shimadzu) with LC-30AD solvent pump, 2020 MS, Sil-30AC autosampler, SPD-M30A UV detector, and CTO-20A column oven; a Corona Veo RS CAD (Thermo Scientific); and a model 8060 CLND (Antek)[10]. Drugs with lower than 80% purity and 20% below expected concentration were excluded.

An Echo 555 acoustic drop ejection (ADE) liquid handler (Labcyte) was fully integrated in the ultra-high-throughput screening uHTS system to dispense DMSO solubilized compounds[11]. Nine-point dose–response curves at 1:3 dilution were generated using ADE as a means of transferring library compounds at ultra-low volume (in nanolitre scale) to achieve direct dilution of compounds. The uHTS system delivered assay-ready daughter plates at ×1,000 concentration. A DMSO backfill step was performed to achieve an equal volume of DMSO in each well. Assay ready drug plates were stored at −80 °C until the day of compound addition and subjected to a single freeze thaw cycle. Use of ADE technology limited the final DMSO concentration in assay plates to 0.1%, which was shown to have a negligible effect on cell growth.

Cell viability assays. Cells were maintained in RPMI-1640, 5% FBS (10% heat-inactivated FBS for suspension lines), and 2 mM glutamine at 37 °C and 5% CO_2 incubators. Cell lines were split at least once before use in cell-based assays. Before plating, adherent lines were trypsinized and suspension lines were re-suspended. Cells were assessed with a Vi-CELL Cell Viability Analyzer (Beckman Coulter); viability of at least 90% for adherent cell lines and 85% for suspensions was required for screening. Cells were plated at previously determined seeding densities. For each adherent cell line, the optimal seeding density intended to achieve 75% confluency at 96 h was assessed by Incucyte (Essen Bioscience). The optimal seeding for suspension and mixed suspension/adherent cell lines was 75% $signal_{max}$ at 96 h, as assessed by CellTiter-Glo.

A Multidrop Combi Reagent Dispenser (Thermo Scientific) was used for plating 25 μl cell suspension into barcoded Falcon 384-well black clear bottom plates. Cell plates were placed in a 37 °C humidified incubator maintained at 5% CO_2 overnight. The next day, copies of the × 1,000 daughter plates were taken out to thaw at room temperature. Room temperature RPMI-1640 without FBS was added to yield ×6 intermediate plates. Drug plates were incubated at room temperature for 1 h. Five microlitres of compound addition was delivered to cell plates using an Oasis integrated Liquid Handling Robot Platform (Dynamic Devices) with Liconic Incubators and automated barcode tracking. After 72 h, cell plates were taken out of the incubator and equilibrated to room temperature for 30 min. Twenty-five microlitres of CellTiter-Glo reagent (Promega) was added using a Biotek Multiflo Microplate Dispenser. Cell lysis was induced by mixing for 30 min on an orbital shaker; plates were then incubated at room temperature for 10 min to stabilize luminescent signal. It is important to note that many of the blood lines required additional time (1–1.5 h) to induce cell lysis on the orbital shaker. Luminescence readout (relative light units; RLU) was performed on a 2104 EnVision Multilabel Plate Reader (PerkinElmer).

Three to four independent biological replicates were produced. Same-plate technical replication was observed to have negligible effect on outcome and therefore omitted.

Cell line authentication and quality control. First, short tandem repeat (STR) profiles were determined for each line using the Promega PowerPlex 16 System. Sixteen loci (fifteen STRs plus Amelogenin for sex determination) were evaluated: D3S1358, TH01, D21S11, D18S51, Penta E, D5S818, D13S317, D7S820, D16S539, CSF1PO, Penta D, AMEL, vWA, D8S1179 and TPOX. Results from gCSI cell lines were compared to external STR profiles of cell lines (when available) to confirm cell line ancestry.

In addition to the initial STR assessment, SNP genotypes were re-checked using a Fluidigm multiplexed assays each time new stocks were expanded for cryopreservation. Resulting SNP profiles were compared to previously generated internal and external data (when available), to reconfirm ancestry. In cases where such data were unavailable or cell line ancestry was questionable, DNA or cell lines were re-purchased. SNPs selection for this purpose was based on minor allele frequency and presence on commercial genotyping platforms. The SNP assay included rs11746396, rs16928965, rs2172614, rs10050093, rs10828176, rs16888998, rs16999576, rs1912640, rs2355988, rs3125842, rs10018359, rs10410468, rs10834627, rs11083145, rs11100847, rs11638893, rs12537, rs1956898, rs2069492, rs10740186, rs12486048, rs13032222, rs1635191, rs17174920, rs2590442, rs2714679, rs2928432, rs2999156, rs10461909, rs11180435, rs1784232, rs3783412, rs10885378, rs1726254, rs2391691, rs3739422, rs10108245, rs1425916, rs1325922, rs1709795, rs1934395, rs2280916, rs2563263, rs10755578, rs1529192, rs2927899, rs2848745 and rs10977980.

All stocks were tested for mycoplasma before and after cells were cryopreserved. Two methods were used to minimize false positive or false negative results: Lonza Mycoalert kit and Stratagene Mycosensor. Cell growth rates and morphology were also monitored for any batch-to-batch changes.

The cell lines MEL-HO, ONCO-DG-1 and KCI-MOH1 were used in our study and are listed on the NCBI misidentified cell line list (http://www.ncbi.nlm.nih.gov/biosample). We chose to maintain them because their utility here depends not on their true identity, but on their common use among gCSI, GDSC and CCLE.

Drug response statistics. gCSI data were processed using the R statistical computing environment (R Core Team, 2013) and a Genentech-developed analysis package, compareDrugScreens, which is available for download.

To reduce noise and the impact of spatial effects, per-dose viability was computed relative to median RLU from untreated wells physically near drugged wells on the plate. To minimize the impact of isolated aberrant wells, a four-parameter log-logistic model relating viability to log-dose was fit by minimizing the mean of residual absolute value (as an alternative to standard least squares). Cases with a high degree of variability across biological replicates were manually flagged and excluded from further analysis. In some cases, grossly aberrant runs were also dropped, resulting in fewer replicates. Fitted viability statistics for retained biological replicates were integrated via maximum likelihood and a hierarchical model that accounted for both within-run noise and cross-run biological variability. Specifically, fitted values and their standard errors for the final, synthesized curve are estimated on a dose by dose basis, and on the log-viability scale. For a given dose, the true viability value for run i is assumed to be subject to biological variation, and is thus modelled as a Gaussian random variable $\mu_i \approx N(\mu_0, \tau^2)$. Given μ_i the viability estimate for run i is modelled as $\nu_i \approx N(\mu_i, \sigma_i^2)$. In practice, noise about fitted curve clearly varies from run to run, so use of run specific σ_i^2 is important. Single-curve fitted viability (ν_i) and its associated estimated standard error (s_i) provide natural plug-in estimates of μ_i and σ_i. These are treated as fixed, and then μ_0 and τ^2 are estimated by numerically optimized maximum likelihood. (See the estMLE.R file in the compareDrugScreens package for additional detail.) To avoid confidence intervals which fall below 0%, standard symmetric confidence intervals are computed on the log-viability scale and then endpoints are exponentiated. Note that because, unconditionally, $\nu_i \approx N(\mu_0, \tau^2 + \sigma_i^2)$, this meta-analysis style approach induces weighted averaging of runs, with weight inversely proportional to within-run noise level.

The reported IC_{50} is the dose at which cross-run estimated inhibition is 50% relative to untreated wells, that is, absolute IC_{50}. The mean viability statistic is the arithmetic average of the fitted viabilities at each tested dose. Mean viability is equivalent to the area under the log-dose/viability curve divided by the total number of tested doses, and is thus on an interpretable percentage scale. Because the mean viability statistic depends on dosing, it is not expected to produce identical quantitative values for projects using different dosing schedule; it should, however, produce correlated results and order responsive lines similarly.

Published CCLE viability statistics differed from those used by gCSI and GDSC in an important respect. CCLE first profiled all cell lines with a high dose of MG132 (a potent cell-permeable proteasome inhibitor). They then defined the minimum attainable viability for a given cell line (nominal '0') as the viability observed upon treatment of that line with MG132. gCSI and GDSC, on the other hand, took the standard approach of equating nominal '0' and true 0—that is, no observed fluorescence. Absolute and MG132-relative viability and IC_{50} statistics are not directly comparable. To enable comparison, we reversed the MG132 adjustment made by CCLE and then refit dose response curves to the CCLE data using the gCSI

© 2016 Macmillan Publishers Limited. All rights reserved

algorithms. (Both the code necessary for the results of this process are provided in the compareDrugScreens package.) All CCLE mean viability and IC_{50} statistics referenced in this manuscript are therefore on absolute rather than MG132-relative scale (Supplementary Table 2). GDSC did not provide per-dose viabilities, so their data could not be reprocessed in this fashion.

Additional investigation of published CCLE dose–response curves identified a small fraction with erroneous reported IC_{50} values; typically, those with per-dose viability values that were either all above or all below the 50% threshold. The recalculation of absolute viabilities described above also corrected this problem.

Assigning sensitive and resistant labels. Specifically active targeted agents were assumed to have no effect on most cell lines, but to inhibit a small fraction of lines dependent on the relevant pathways, to varying degrees. To model this behaviour, two component Gaussian mixture distributions were fit to untruncated gCSI and CCLE mean viability data for crizotinib, erlotinib and lapatinib, using mclust (http://bioconductor.org). Although Extended Data Fig. 2a gives bivariate fits for illustration purposes, final sensitive/resistant thresholds were assigned one study at a time. Variance for the two components was not required to be equal. Cell lines more than two standard deviations below the fitted resistant group mean were called sensitive. The fitted sensitive group was more dispersed since, unlike the resistant group, variation here reflects both assay noise and true variation in degree of sensitivity to the drug; to accommodate this, lines more than one standard deviation above the fitted sensitive group mean were called resistant. Lines that did not meet either of these criteria we assigned no-call. Extreme outliers (>1.05 for gCSI and >1.10 for CCLE) were excluded from fitting, but not classification, for crizotinib, erlotinib and lapatinib as the two-component fit is susceptible to assigning these rare outliers their own distribution.

This strategy could not be applied to the GDSC mean viability results owing to truncation of published statistics, which precludes the use of Gaussian distributions. Instead, we used the same cutoffs for gCSI as computed above, counted lines called sensitive or resistant by gCSI among those in the gCSI/GDSC overlap set, and then assigned the same number of sensitive or resistant labels to those cell lines with most extreme GDSC mean viability scores. GDSC truncation of mean viability values at 1 often resulted in ties; these were resolved at random in order to select an appropriate number of resistant lines.

Broadly active agents were assumed to affect most or all lines, although to widely varying degrees. This makes use of mixture distributions inappropriate. For these compounds, we simply split cell lines into three even groups and assigned sensitive, no-call and resistant labels.

When assigning labels for data from a given institution, all available lines from that institution were used, even those not screened by other groups.

Expected biomarkers and expected responder cell lines. For Fig. 2 and related analyses, we selected sets of cell lines (termed 'expected responders') with the clearest evidence of genomic features previously shown to correlate with response to each of six drugs. These lines were selected using the gCSI feature-standardized genomic data matrix, so all cutoffs for quantitative variables are in standard (z-score) units. For crizotinib, the following features and cutoffs were used: *MET* amplification (copy number z-score >3.0) or evidence of *EML4*/*ALK* translocation. Further, the cell line OE33 lacked copy number data but is *MET* amplified[12]; the cell line NB-1 expresses an *AML4*/*ALK* fusion transcript[13] and SR-786 has a *NPM1*/*ALK* fusion[13]. Both were included as expected responders to crizotinib. For erlotinib, we selected lines with *EGFR* hotspot mutation (Supplementary Table 6) or high *EGFR* expression; for irinotecan, lines with high *SLFN11* expression[1,14]; for lapatinib, lines with *EGFR* hotspot mutation, high *EGFR* expression or high *ERBB2* expression. The biomarker for paclitaxel was low *ABCB1* expression[15]. Finally, the biomarkers for PD0325901 were a positive value for the MAPK metafeature or high *SPRY2* expression. Gene expression cutoffs were selected based on inspection of expression density plots (not shown). A cutoff of 2 was used to isolate the upper tail of unimodal distributions (*EGFR*, *ERBB2*, *SPRY2*), while a cutoff was placed in the gap for bimodal distributions (-1 for *ABCB1*, 0 for *SLFN11*).

Selecting lines for follow-up screening. Twenty-four lines were selected for the factorial assessment of viability assay, growth media, and cell seeding strategy. Because the SYTO 60 protocol is challenging to implement with cells in suspension[2], we restricted attention to adherent lines. Twelve specific features of the 177 cell lines available in all three studies were ranked. These features quantified differences in mean viability or IC_{50} among the three projects, availability of response data for cases where many lines were missing values, and relative seeding density in the gCSI primary screen. Cell lines with the highest average rank were selected.

Follow-up screen. Cells were maintained in RPMI-1640, 5% FBS and 2 mM glutamine. Cells were assessed with a Vi-CELL Cell Viability Analyzer (Beckman Coulter); viability of at least 90% for all cell lines was required for screening. Cells were plated at optimal seeding density to a previously determined, line-specific level intended to achieve 75% confluency at 96 h using an Incucyte (Essen Bioscience). In addition, the same cell lines were also plated at a fix seeding density of 1,020 cells per well. To address whether serum concentration had an effect on drug response, cells were plated at 5% and 10% respectively. Multidrop Combi Reagent Dispenser (Thermo Scientific) was used for plating cells into barcoded Falcon 384-well black clear bottom plates. Cell plates were placed in a 37 °C humidified incubator maintained at 5% CO_2 overnight. A 9-point dose dilution was performed using an Echo 550 (Labcyte). The next day, compound addition was done using a Bravo Automated Liquid Handling Platform (Agilent Technologies). After 72 h, 25 μl CTG reagent (Promega, G7573) was added using a Biotek Multiflo Microplate Dispenser. Cell lysis was induced by mixing for 2 min on an orbital shaker; plates were then incubated at room temperature for 10 min to stabilize luminescent signal. Luminescence readout (RLU) was measured on a 2104 EnVision Multilabel Plate Reader (PerkinElmer). An identical set of assay plates was performed using SYTO 60 Red Fluorescent Nucleic Acid Stain method (Life Technologies, S11342) to compare different types of endpoint readout. Cells were fixed for 30 min using 4% formaldehyde prepared in PBS. Formaldehyde was aspirated and plates were washed 3 times using automatic liquid handling. Cell nuclei were stained using the SYTO 60 at 1 μM, 50 μl per well (384-well plate). Plates were incubated in the dark for 1 h at room temperature. Dye was aspirated and plates were washed three times using automatic liquid handling. Fluorescence intensity is measured (plate bottom measurement) by EnVision at 652 nm/678 nm. Three independent biological replicates were produced; same-plate technical replication was observed to have negligible effect on outcome and therefore omitted.

Gene expression. Expression levels were quantified by RNASeq via Illumina sequencing (75-bp paired-end reads) in one sample per cell line. Reads were filtered for quality and for rRNA contamination. These reads were aligned to the genome (GRCh37.1) using GSNAP (version 2013-03-31)[16] with the following options: -M 2 -n 10 -B 2 -i 1 -N 1 -w 200000 -E 1-pairmax-rna = 200000 -clip-overlap. An average of 51.3 million uniquely aligned reads with concordant read pairs were available per sample. Read counts for each transcript were processed with the DESeq variance stabilizing transform (VST) method[17] to obtain normalized expression levels. The resulting expression measurements are referred to as 'VST values'.

DNA copy number and loss of heterozygosity. Illumina Human Omni 2.5-4 or Omni 2.5-8 arrays, one per cell line, were processed with a modified version of the PICNIC (ref. 18) algorithm, as published recently[19]. In short, the modifications to PICNIC increase the accuracy of global ploidy (and normal cell contamination) estimation and make the algorithm compatible with Illumina arrays. These modifications also allow for accurate total copy number estimates and dramatically reduce over-segmentation. After PICNIC determines the sample ploidy and performs segmentation with its hidden Markov model (HMM) to determine allele-specific copy number, we added a final segmentation with cghFLasso[20]. This addition accounts for the over-segmentation of PICNIC in regions where point copy number estimates indicate non-integer total copy number. This segmented total copy number was averaged for each gene using all SNP intervals that include a portion of that gene. We used the absolute copy number scale (where 2 is normal for an autosome) rather than the traditional relative $\log_2$ absolute/ploidy copy number. Loss of heterozygosity (LOH) was assessed using the lesser of the two allele-specific copy number values (LACN), from the PICNIC HMM, using the same SNPs. Genes with an average LACN value <1 were considered to have LOH. Absolute total copy number values and binarized LOH values were included in the model for every gene.

Mutations. DNA mutations were assessed by Illumina exome sequencing (75-bp paired-end reads) in one sample per cell line (manuscript in preparation). GSNAP (version 2013-03-31)[16] was used to align reads to the genome (GRCh37.1) with the following options: -M 2 -n 10 -B 2 -i 1-pairmax-dna = 1000-terminal-threshold = 1000-gmap-mode = none-clip-overlap. Duplicate alignments were filtered using *PICARD* tools (http://broadinstitute.github.io/picard). An average of 64.7 million uniquely aligned reads with concordant read pairs were available per sample. These aligned reads were used to assess genomic SNV and indel variants using the VariantTools (http://bioconductor.org) package (version 1.4.5) and the following cutoffs: a minimum of 2 reads supporting the variant and at least 5% variant frequency.

We assessed our cell lines for mutations in 235 hotspot codons in specific RefSeq transcripts for 68 genes (Supplementary Table 6). As mutation consequences are evaluated with respect to a specific transcript, we ensured that the selected transcripts indicated the expected wild type amino acids at the hotspot codons. Cell lines with any missense mutation within a hotspot codon were counted as mutant for that hotspot. A cell line lacking a variant at a given hotspot codon was counted as wild type if all three bases in the codon had sufficient coverage to detect a variant, with 80% power, given the VariantTools cutoffs above, and otherwise was considered a no-call (NA).

Hotspots were also combined, by gene, such that a cell line with any hotspot mutation in a given gene was considered positive for that hotspot metafeature.

© 2016 Macmillan Publishers Limited. All rights reserved

For example, a cell line with a *PIK3CA* missense mutation at amino acid 542, 545 or 1047 was considered positive for the *PIK3CA* hotspot metafeature.

To address tumour suppressor genes or previously uncharacterized mutations in oncogenes, three levels of mutations were considered for each coding position in each gene. First, mutation consequences were annotated using the Ensembl Variant Effect Predictor (VEP)[21]. VEP implements a controlled vocabulary, and VEP vocabulary terms in the following set were considered to always be function altering mutations (and coded as mut): TRANSCRIPT_ABLATION, ESSENTIAL_SPLICE_SITE, STOP_GAINED, FRAMESHIFT_CODING, STOP_LOST. Missense mutations that had previously been observed one or more times in patient tumours (COSMIC, version v63_300113)[22] were also considered to be function-altering mutations (mut). Second, a larger mutp class was formed by combining mut variants with non-COSMIC missense variants assessed as 'damaging' by the Condel tool[23]. Finally, variants that met the mutp criteria and were also within a segment of LOH were coded as mutpl. For each of these three categories of mutation, a sample was called wild type for a given gene if it had no variants and also had sufficient coverage at $\geq$50% of coding nucleotides; genes with no variants but insufficient coverage were considered NA.

Finally, we also formed three pathway metafeatures by combining mutations. The PI3K metafeature was positive for lines with a *PIK3CA* hotspot mutation, *PTEN* LOH, or *PTEN* loss. The DNA damage metafeature was positive for any cell line with a mutp variant in *MLH1*, *MLH3*, *KMT2D* (also known as *MLL2*), *MSH2*, *MSH3*, *MSH6*, *PMS1* or *PMS2*. The MAPK metafeature was positive for any cell line with a hotspot mutation in *BRAF*, *HRAS*, *KRAS*, or *NRAS*.

GDSC and CCLE data. For CCLE, DNA copy number and expression were downloaded from GEO (accession number GSE36139). Mutation and cell line description data were downloaded from the CCLE website (http://www.broadinstitute.org/ccle/home). Drug and cell line names were matched across data sets. (Note that the cell line 'T.T' was renamed 'tdott' to differentiate from cell line 'TT'.) As described above, the CCLE IC_{50} and mean viability values were re-calculated using the provided per-dose viability values from ref. 24 (see Supplementary Table 2).

The GDSC mean viability and IC_{50} values (gdsc_manova_output_w2.csv) were downloaded from ftp://ftp.sanger.ac.uk/pub4/cancerrxgene/releases/release-2.0 as in ref. 3. Note that GDSC mean viability statistics were not included in the original publication. Because only summary statistics were provided, GDSC mean viability values were used as is. The GDSC extrapolated IC_{50} values, on the other hand, were truncated to the limits of the tested dose range for consistency with gCSI and CCLE data.

Predictive modelling. In preparation for predictive modelling, several cell line features were gathered: tissue of origin; growth rate as assessed by seeding density (the number of cells per assay control well required to reach 75% confluency at 72 h); and doubling time.

All sample descriptors were combined into one matrix for predictive modelling. Expression VST values, DNA copy number, seeding density and growth rate were expressed as continuous values. Mutation status for individual hotspots, for gene hotspot and pathway metafeatures, and for the three categories of function-altering mutation were each individually coded as binary variables. Categorical variables, such as tissue type, were coded with one binary dummy variable per factor level. For example, the value '1' for the 'isLung' feature indicates that a cell line is derived from lung cancer. All features, including the binary features, were standardized before imputation and predictive modelling.

Missing data, owing to a lack of genomic data or insufficient sequencing coverage, were imputed using glmnet[4]. Missing values were initially replaced with the average value for their column (feature). Missing data were then imputed, one feature at a time, in a serial fashion, by treating one column as the response variable and all other feature columns as dependent variables. Training was performed using all samples with data for the response variable and values were then predicted for the samples previously missing data.

Classification and feature selection were performed using glmnet[4] with binarized or continuous valued mean viability as the response variable. We selected an α value of 0.5 to increase the richness of the selected features and to increase interpretability. We selected the minimum lambda value within one standard deviation of the absolute minimum value under cross validation. This cross validation was performed with 25 randomly selected hold-out groups consisting of 20% of the samples. Hold-out groups were selected using the 'caret' (http://bioconductor.org) tool to select sample groups with sensitive/resistant proportions or continuous response variable distributions similar to that of the full set of samples. Following cross validation and lambda selection, a glmnet run with all samples was performed to select potential biomarkers. All features with a non-zero coefficient under the selected lambda were used for further analyses.

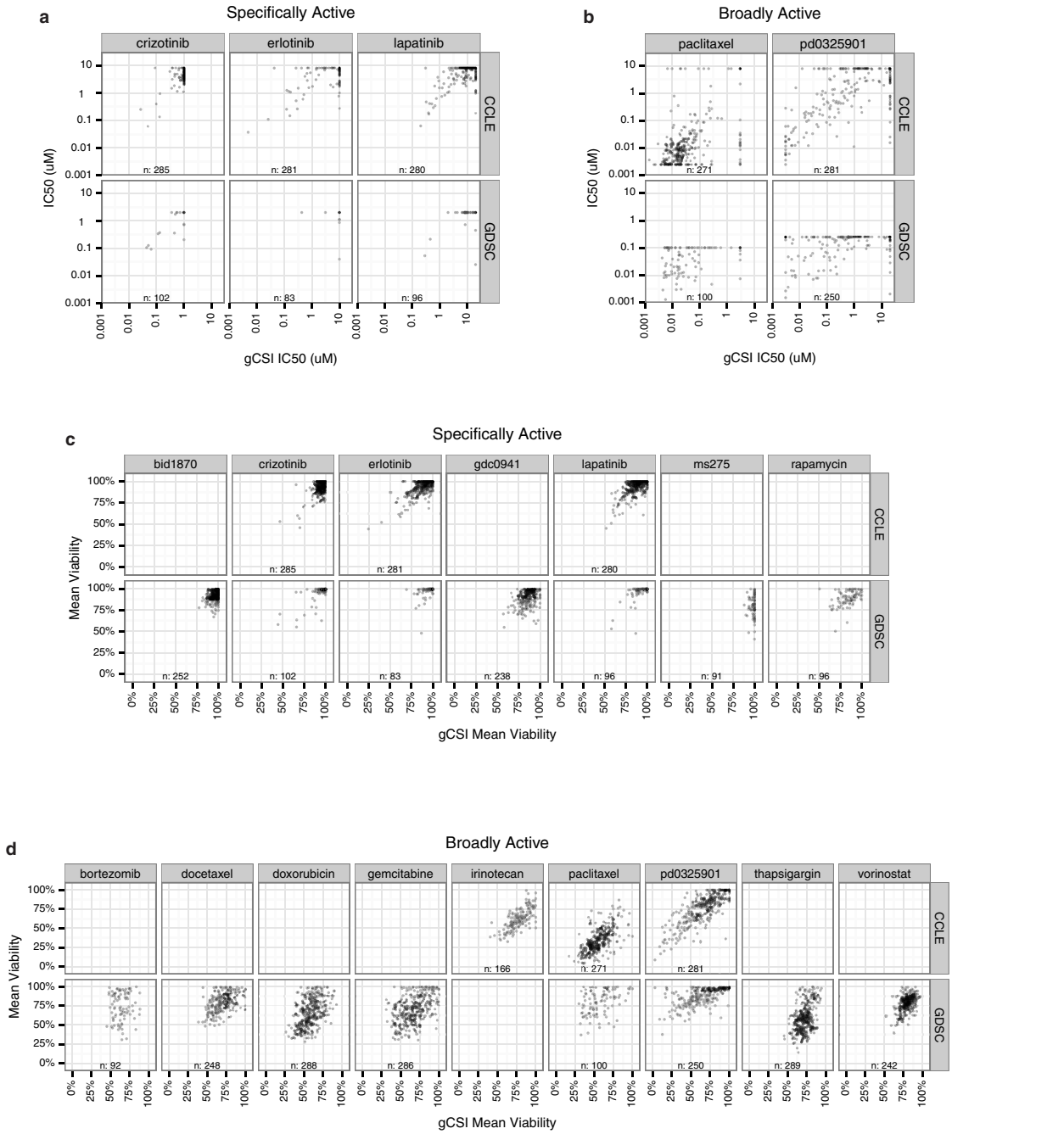
Owing to large differences in genomic feature naming and availability, it was necessary to use only one genomics data set at time when scoring genomic features for correlation with drug response and selecting candidate biomarkers. gCSI and CCLE mean viability data were each compared to the gCSI genomics data (Fig. 3 and Extended Data Fig. 5a), and then separately to the CCLE genomics data (Extended Data Figs 5b–d). A similar analysis was performed for gCSI versus GDSC discretized mean viability statistics, relative to gCSI genomics data (Extended Data Fig. 7). Owing to sporadically missing drug response data and differences in the sets of lines for which gCSI and CCLE produced genomics data, the number of cell lines used in each comparison varied (Supplementary Table 4). For Fig. 3 and Extended Data Figs 5 and 6, only cell lines with gCSI and CCLE drug response data and genomic data were used. For Extended Data Fig. 7, only samples available in the gCSI and GDSC drug data and gCSI genomics data were used.

Univariate analysis of drug response and genomics. Drug response data and standardized genomics data used for predictive modelling (above) were also analysed one genomic feature at a time. Continuous drug response data was regressed on each genomic feature using standard least squares. Logistic regression was used for discretized drug response data. Correction for multiple testing was performed using the Benjamini and Hochberg method as implemented in the R package multtest (http://bioconductor.org).

Code and data availability. All source code used in this study has been made available in the R computer language, in a fully documented software and data package: compareDrugScreens. This package is freely available under the Creative Commons 3.0 license and can be downloaded from http://research-pub.gene.com/gCSI-cellline-data. All gCSI, CCLE and GDSC drug and genomic data used in this study are also included in this package.

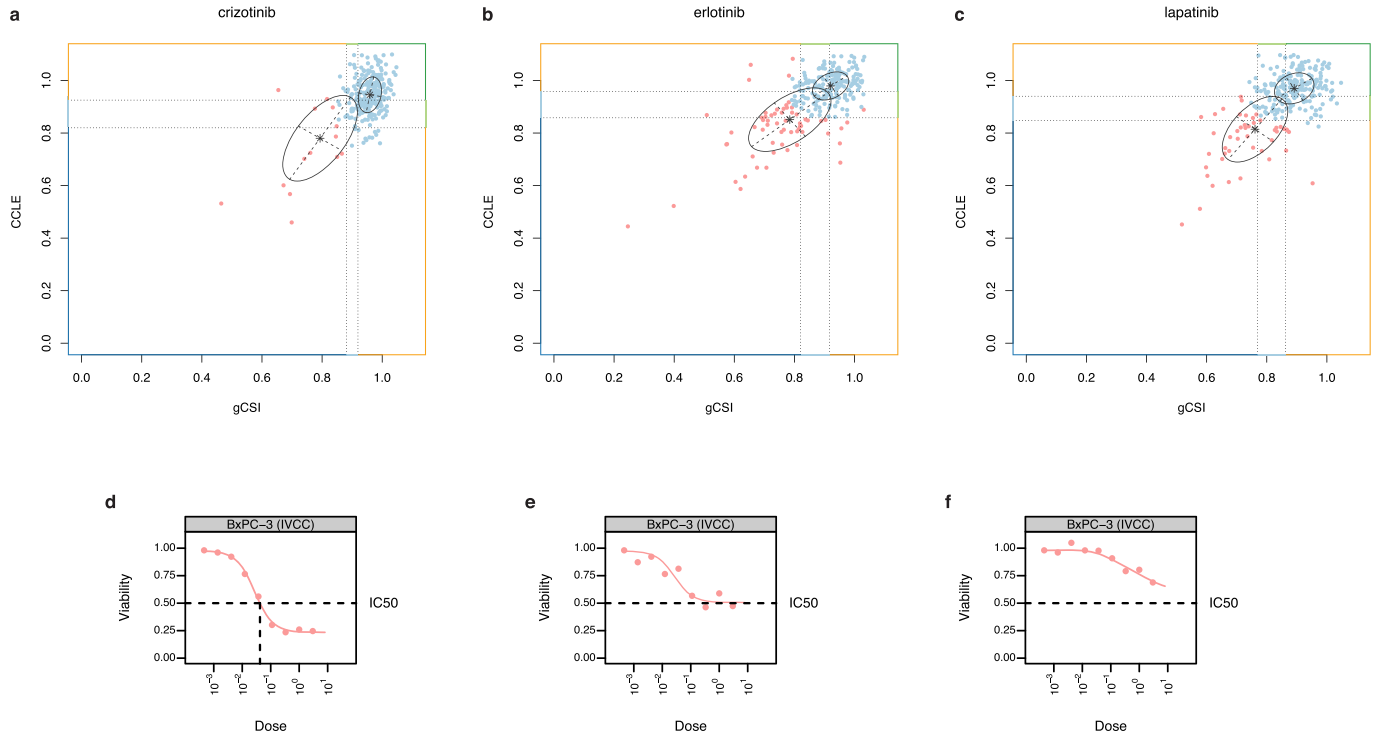
9. Yu, M. *et al.* A resource for cell line authentication, annotation and quality control. *Nature* **520,** 307–311 (2015).
10. Jiang, Y., Hascall, D., Li, D. & Pease, J. H. A charged aerosol detector/chemiluminescent nitrogen detector/liquid chromatography/mass spectrometry system for regular and fragment compound analysis in drug discovery. *J. Chromatogr. A* **1411,** 63–68 (2015).
11. Dawes, T. D. *et al.* Compound transfer by acoustic droplet ejection promotes quality and efficiency in ultra-high-throughput screening campaigns. *J. Lab. Autom.* **21,** 64–75 (2016).
12. McDermott, U. *et al.* Identification of genotype-correlated sensitivity to selective kinase inhibitors by using high-throughput tumor cell line profiling. *Proc. Natl Acad. Sci. USA* **104,** 19936–19941 (2007).
13. Klijn, C. *et al.* A comprehensive transcriptional portrait of human cancer cell lines. *Nature Biotechnol.* **33,** 306–312 (2015).
14. Zoppoli, G. *et al.* Putative DNA/RNA helicase Schlafen-11 (SLFN11) sensitizes cancer cells to DNA-damaging agents. *Proc. Natl Acad. Sci. USA* **109,** 15030–15035 (2012).
15. Duan, Z., Lamendola, D. E., Duan, Y., Yusuf, R. Z. & Seiden, M. V. Description of paclitaxel resistance-associated genes in ovarian and breast cancer cell lines. *Cancer Chemother. Pharmacol.* **55,** 277–285 (2005).
16. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26,** 873–881 (2010).
17. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).
18. Greenman, C. D. *et al.* PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. *Biostatistics* **11,** 164–175 (2010).
19. Rudin, C. M. *et al.* Comprehensive genomic analysis identifies *SOX2* as a frequently amplified gene in small-cell lung cancer. *Nature Genet.* **44,** 1111–1116 (2012).
20. Tibshirani, R. & Wang, P. Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics* **9,** 18–29 (2008).
21. Yourshaw, M., Taylor, S. P., Rao, A. R., Martín, M. G. & Nelson, S. F. Rich annotation of DNA sequencing variants by leveraging the Ensembl Variant Effect Predictor with plugins. *Brief. Bioinform.* **16,** 255–264 (2015).
22. Forbes, S. A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39,** D945–D950 (2011).
23. González-Pérez, A. & López-Bigas, N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am. J. Hum. Genet.* **88,** 440–449 (2011).
24. Barretina, J. *et al.* Addendum: The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **492,** 290 (2012).
25. Sigoillot, F. D. & King, R. W. Vigilance and validation: Keys to success in RNAi screening. *ACS Chem. Biol.* **6,** 47–60 (2011).
26. Fraley, C. & Raftery, A. E. Model-based clustering, discriminant analysis and density estimation. *J. Amer. Statist. Assoc.* **97,** 611–631 (2002).

© 2016 Macmillan Publishers Limited. All rights reserved



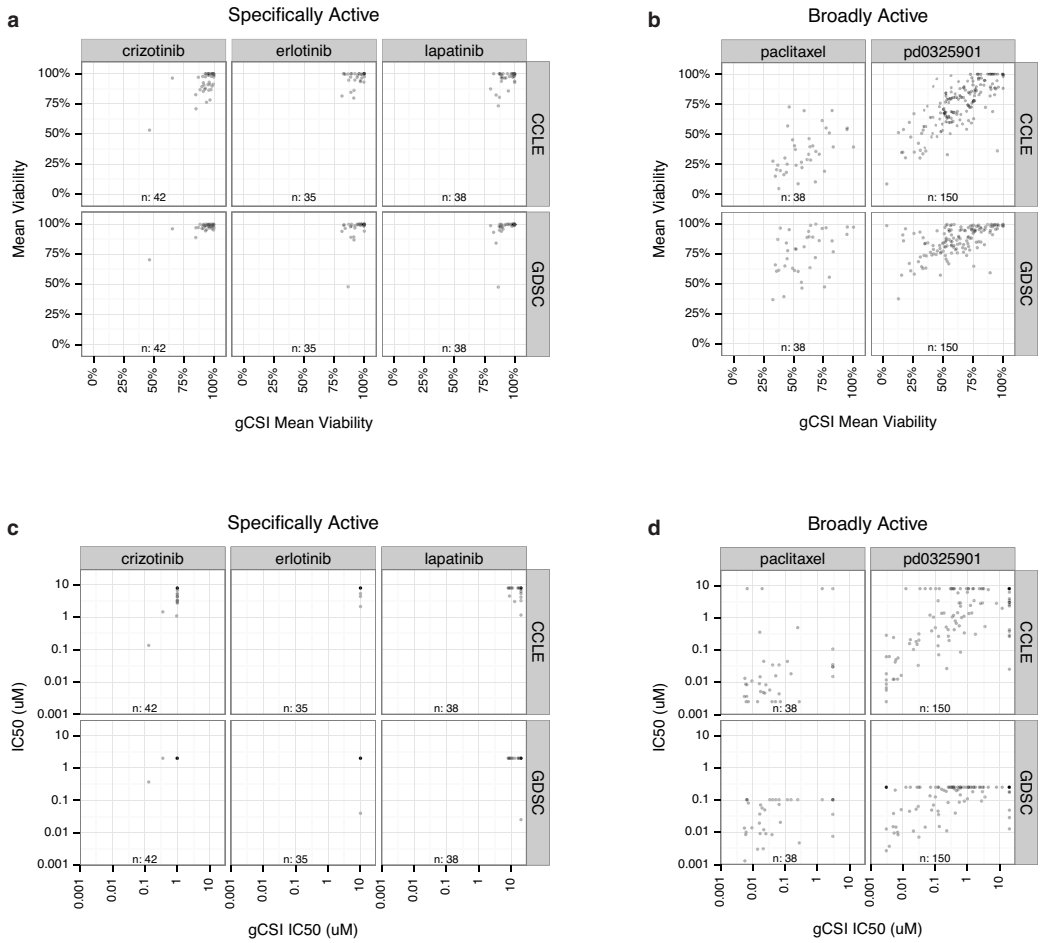
Extended Data Figure 1 | Comparison of gCSI, CCLE and GDSC mean viability response statistics: all pairwise comparisons. a, **b**, IC_{50} values (the drug concentration estimated to produce a 50% reduction in viability relative to untreated cells) for the cell lines in Fig. 1. Because extrapolation beyond tested doses can produce statistically unstable results (Extended Data Fig. 2b), IC_{50} values are truncated to the tested dose range (which varies by drug and project). **c**, **d**, Mean viability (equivalent to area under the log-dose/response curve) is plotted for all drugs that gCSI had in common with one or both other studies. 0% corresponds to complete growth inhibition at all doses; 100% corresponds to viability equal, on average, to untreated control wells. Values are capped at a maximum value of 100% for consistency with GDSC published data. gCSI mean viability and IC_{50} values were derived from three biological replicates (Methods).

© 2016 Macmillan Publishers Limited. All rights reserved



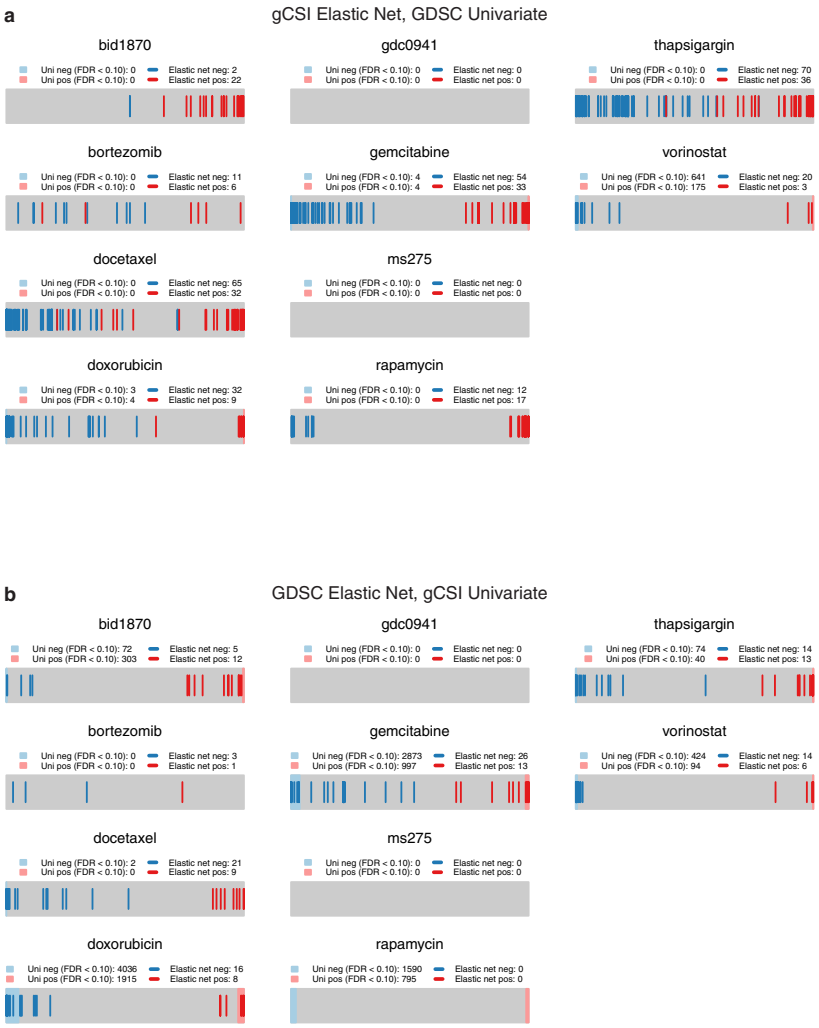
Extended Data Figure 2 | Assignment of sensitive and resistant labels for specifically active agents. a–c, Untruncated mean viability statistics from CCLE and gCSI are plotted for three drugs with specific activity. Two-component Gaussian mixture model fits to these values were performed to classify non-responder and responder distributions for CCLE and gCSI separately[26]. Non-responder mean viability statistics were scattered around 100%, providing a good indication of assay noise levels. Point colour indicates hard sensitive (red) versus resistant (blue) thresholding based on likelihood ratio—a strategy that ignores uncertainty of classification. Dotted lines, on the other hand, indicate the final thresholds selected by the method used for Fig. 2; this method quantifies uncertainty and, if it is too high, assigns a no-call label to the cell line (Methods). Ellipses indicate the covariance from a two-dimensional, two-component fit and are for illustrative purposes only; response labels were assigned on the basis of individual fits for gCSI and CCLE separately, using all available cell lines for each data set. **d–f**, Hypothetical log-dose/viability curves illustrating statistical challenges presented by IC_{50} response metric. In small screens, it is often feasible to adjust dosing scale so that a 50% inhibition (relative to vehicle-treated wells at the same time point) is cleanly achieved. For large-scale screening, on the other hand, it is common for many cell lines to fail to achieve this level of inhibition at any tested dose, and/or to fail to demonstrate a lower asymptote in their dose–response curve. The mean viability statistic, on the other hand, requires no extrapolation beyond tested doses. **d**, A hypothetical dose–response curve with characteristics ideal for relative or absolute IC_{50} computation. Tested doses yield viability values well above and well below 50%. Furthermore, both upper and lower asymptotes can be precisely estimated, enabling use of relative IC_{50}, if desired (relative IC_{50} is computed after first rescaling to the observed viability dynamic range. In the present work, however, all three studies use absolute IC_{50}). **e**, A hypothetical dose–response curve that yields precise upper and lower asymptote estimates but a highly imprecise absolute IC_{50} estimate: doses over a range exceeding two orders of magnitude are all consistent with 50% inhibition, and raw data are too noisy for further refinement. Use of such IC_{50} values, which can be frequently observed in gCSI, CCLE and GDSC results, may complicate cross-study comparisons. **f**, A hypothetical dose–response curve that requires potentially imprecise extrapolation for estimation of both absolute IC_{50} and lower asymptote. We advise against use of either absolute or relative IC_{50} for such curves, which are again common in all three studies under consideration. Finally, also note that in settings such as this, where low doses provide no information and inhibition occurs only at the highest tested doses, any response statistic is forced to rank cell lines on the basis of just a subset of the data generated. As a consequence, results for such drugs may exhibit reduced concordance regardless of analysis method.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 3 | Comparison of gCSI, CCLE and GDSC primary drug response statistics: only cell lines assayed by all three projects. Equivalent to Fig. 1, but restricted to cell lines in the three-way gCSI, CCLE and GDSC intersection. Figure 1 shows a significant number of cell lines interrogated by gCSI and CCLE that are sensitive to the three specifically active drugs. Of these, few (if any) are visible here, because few were included in the gCSI–GDSC intersection. **a**–**d**, Mean viability statistics (**a**, **b**), and IC_{50} (**c**, **d**). Both are truncated as for Fig. 1 and Extended Data Fig. 1. gCSI mean viability and IC_{50} values were derived from three biological replicates (Methods).

© 2016 Macmillan Publishers Limited. All rights reserved



c **Agreement rate for doubly-called cell lines screened only by gCSI and GDSC**

		GDSC					
			Agree				
Type	Drug	Called (%)	Sensitive	Resistant	Disagree	Agree (%)	
specific	**bid1870**	30%	13	40	23	**70%**	**7.7E-03**
broad	**bortezomib**	52%	13	13	22	**54%**	0.57
broad	**docetaxel**	48%	46	49	25	**79%**	**1.6E-10**
broad	**doxorubicin**	47%	57	53	26	**81%**	**9.6E-14**
specific	**gdc0941**	47%	3	103	5	**95%**	**8.6E-04**
broad	**gemcitabine**	44%	42	59	24	**81%**	**4.2E-12**
specific	**ms275**	62%	2	33	21	**63%**	0.71
specific	**rapamycin**	44%	9	26	7	**83%**	**1.8E-04**
broad	**thapsigargin**	46%	41	56	36	**73%**	**2.2E-07**
broad	**vorinostat**	50%	45	51	26	**79%**	**2.2E-10**

Extended Data Figure 4 | gCSI–GDSC concordance for drugs common to only gCSI and GDSC. a, Genomic features were ranked by direction and strength of their association with GDSC mean viability (from strongest negative to strongest positive). Shaded regions indicate features with significant univariate association (FDR-adjusted $P < 0.10$). The univariate ranks of features selected by the elastic net using gCSI drug response are indicated with vertical bars, also coloured by direction of association. **b**, As for panel **a**, but applying the elastic net to GDSC drug response data and comparing results to univariate statistics obtained from gCSI drug response. **c**, As in Table 1, lines called sensitive or resistant by both gCSI and GDSC were assessed for concordant classification.

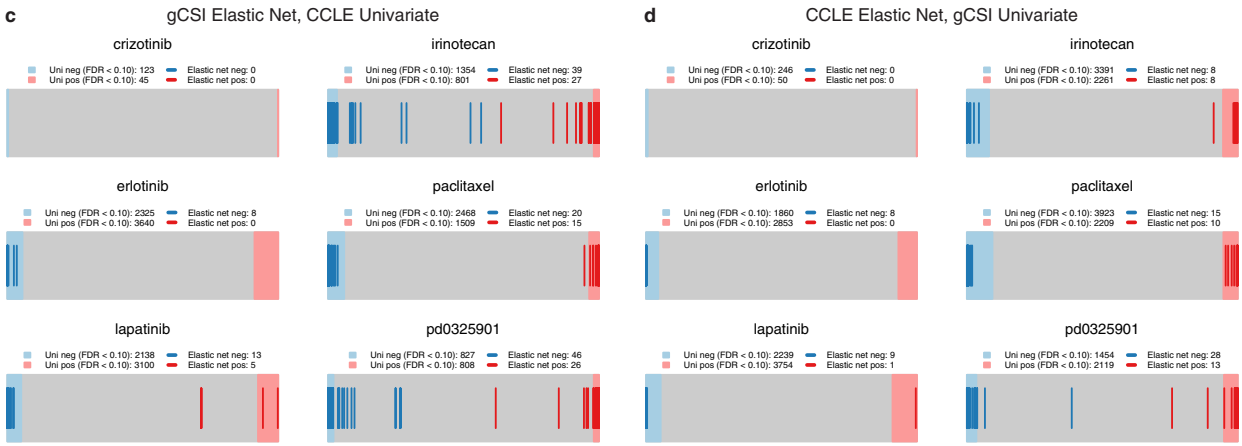
© 2016 Macmillan Publishers Limited. All rights reserved

a **Features identified by elastic net for gCSI and CCLE mean viability data (gCSI genomics)**

Drug	Target	Expected Biomarkers	Selected positive control features		Total features		Shared Genes
			gCSI	CCLE	gCSI	CCLE	
crizotinib	ALK, MET	MET, HGF expression, MET copy number	ST7-AS1		18	0	
erlotinib	EGFR	EGFR mutation, expression, or copy number	EGFR	EGFR	9	13	CDH3, EGFR, ITGB6, LOC100507500, LOC100652762
lapatinib	ERBB2	EGFR or ERBB2 copy number or expression, EGFR mutation	ERBB2	ERBB2	18	8	ERBB2, MIR4728
irinotecan	Topoisomerase	SLFN11 expression	SLFN11	SLFN11	75	20	SLFN11, SQSTM1
paclitaxel	Tubulin	ABCB1 expression	ABCB1	ABCB1	217	36	ABCB1, ARF4, BCL2L1, C1orf9, LOC100506325, OR6A2, SLC35F5, SSRP1, TRPC4AP
pd0325901	MEK	SPRY, DUSP, ETV4 expression, RAS/RAF mutation	SPRY2, MAPK metafeature, ETV4, DUSP6	MAPK metafeature, SPRY2, DUSP6	37	29	DNAJA4, DUSP6, ERC2, GUF1, HSD17B11, HSD17B6, MAPK metafeature, RNF125, SPRY2, TMEM63C

b **Features identified by elastic net for gCSI and CCLE mean viability data (CCLE genomics)**

Drug	Target	Expected Biomarkers	Selected positive control features		Total features		Shared Genes
			gCSI	CCLE	gCSI	CCLE	
crizotinib	ALK, MET	MET, HGF expression, MET copy number			0	0	
erlotinib	EGFR	EGFR mutation, expression, or copy number	EGFR	EGFR	8	8	CDH3, EGFR, KLF5, MAL2
lapatinib	ERBB2	EGFR or ERBB2 copy number or expression, EGFR mutation	ERBB2	ERBB2	18	10	ERBB2, KLF5, MAL2, RHOV
irinotecan	Topoisomerase	SLFN11 expression	SLFN11	SLFN11	66	16	ATP6V0E1, SLFN11
paclitaxel	Tubulin	ABCB1 expression	ABCB1	ABCB1	35	25	ABCB1, PPP1R9A, SSRP1, TMEM139
pd0325901	MEK	SPRY, DUSP, ETV4 expression, RAS/RAF mutation	SPRY2, ETV4, BRAF, KRAS	SPRY2, ETV4, DUSP6, BRAF	72	41	BRAF, CCDC107, ETV4, HIVEP1, HSD17B6, LTBP1, MAGEE1, NPAS2, RNF125, SFXN3, SPRY2, SSFA2

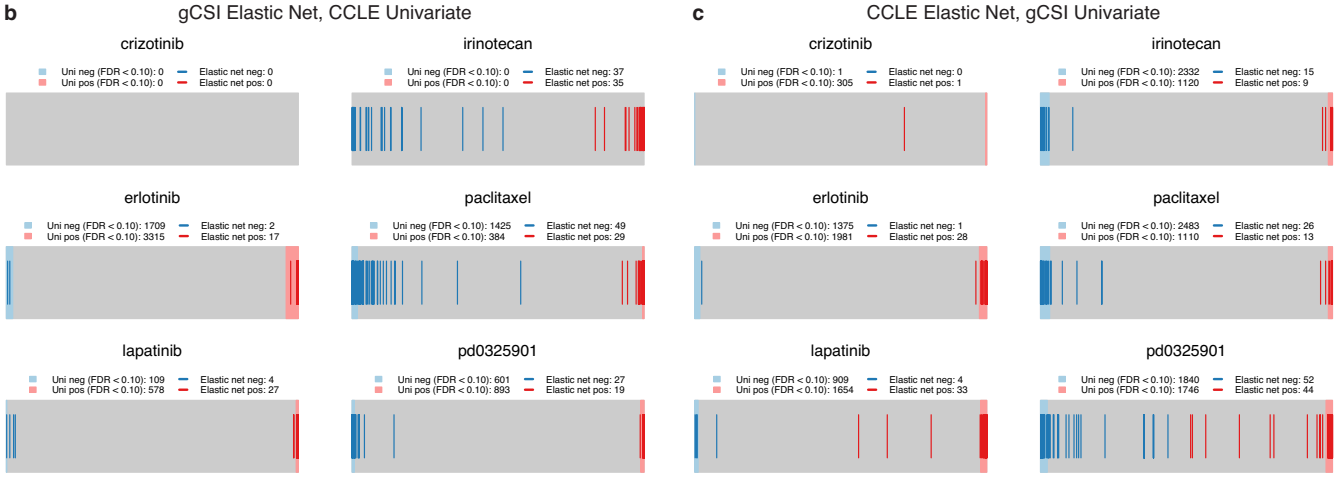


Extended Data Figure 5 | Consistent biomarker scoring for gCSI and CCLE drug response data. The elastic net algorithm was applied to identify short lists of candidate biomarkers for drug response as assessed by either gCSI or CCLE mean viability values, using gCSI genomic data. Supplementary Table 4 describes the number of cell lines available for each comparison. **a**, Description of biomarkers expected to associate with response for each of the six drugs tested by both gCSI and CCLE; recovery of expected biomarkers; total selected feature count; and overlap between gCSI and CCLE selected feature sets. Asterisk denotes that *ST7-AS1* is adjacent to *MET* and *ST7-AS1* expression correlates with *MET* copy number (Pearson correlation $r = 0.47$, $P < 2.2 \times 10^{-16}$, $n = 534$). The gCSI genomics set does not include translocations, so *ALK* translocation could not be checked directly. **b**, Equivalent to panel **a**, but using CCLE rather than gCSI genomics data. **c**, **d**, Equivalent to Fig. 3a, b, but using CCLE rather than gCSI genomics data.

© 2016 Macmillan Publishers Limited. All rights reserved

a **Features identified by elastic net for gCSI and CCLE discretized mean viability data (gCSI genomics)**

			Selected positive control features		Total features		
Drug	Target	Expected Biomarkers	gCSI	CCLE	gCSI	CCLE	Shared Genes
crizotinib	ALK, MET	MET, HGF expression, MET copy number			0	1	
erlotinib	EGFR	EGFR mutation, expression, or copy number		EGFR	19	29	ITGB6, KLF5, LOC100652762
lapatinib	ERBB2	EGFR or ERBB2 copy number or expression, EGFR mutation	ERBB2	ERBB2	31	37	ABCA12, DSG3, EPN3, ERBB2, FAM83H, LOC728643, MAL2, MIR4728, RHOV, SERPINB13, SFN, SPRR3
irinotecan	Topoisomerase	SLFN11 expression	SLFN11	SLFN11	72	24	EFCAB4B, LOC100128816, LRRC34, SLFN11
paclitaxel	Tubulin	ABCB1 expression	ABCB1		78	39	EXOSC2, RIOK1, TIMM8A, TMEM139, TRPC4AP
pd0325901	MEK	SPRY, DUSP, ETV4 expression, RAS/RAF mutation	MAPK metafeature, SPRY2, ETV4, DUSP6	MAPK metafeature, SPRY2, DUSP6	46	96	ADAMTSL3, DUSP6, FAM55C, GUF1, HSD17B11, LOC100505839, MAPK metafeature, RNF125, SH3TC2, SPRY2, ZCCHC9

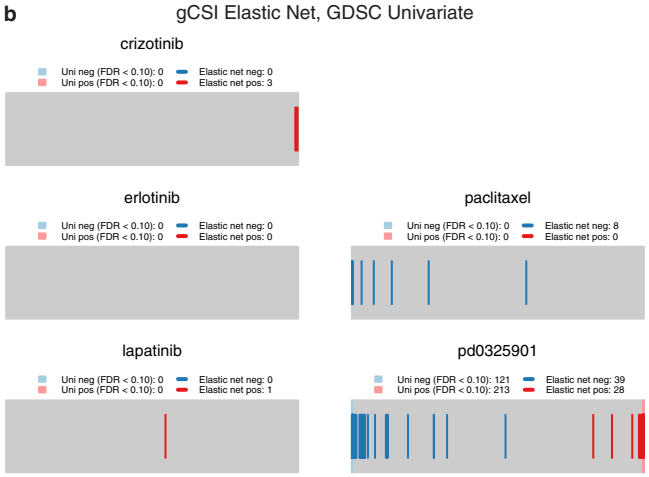


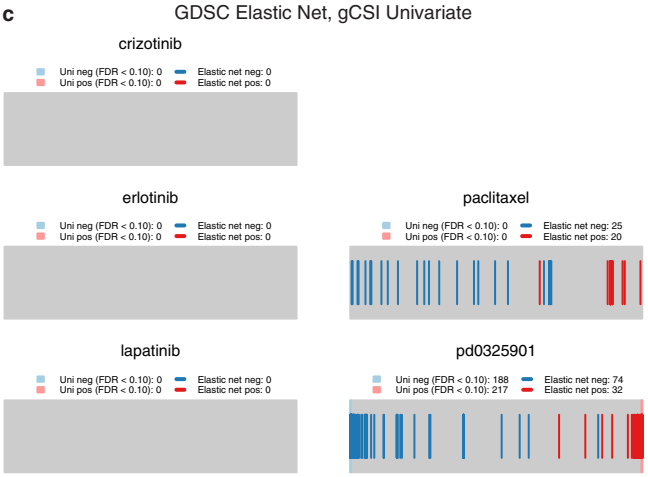
Extended Data Figure 6 | Features identified as response biomarkers using gCSI and CCLE discretized mean viability drug response data versus gCSI genomics data. Equivalent to Fig. 3 and Extended Data Fig. 5a, but using discretized sensitive/resistant labels rather than continuous drug response statistics (lines assigned a no-call label were ignored for this analysis). Recovery of expected biomarkers and scoring of features selected by elastic net in the opposing project's univariate analysis are both comparable to results obtained from continuous mean viability statistics (Fig. 3 and Extended Data Fig. 5a). **a**, Description of biomarkers expected to associate with response for each of the six drugs tested by both gCSI and CCLE; recovery of expected biomarkers; total selected feature count; and overlap between gCSI and CCLE selected feature sets. **b**, Genomic features were ranked by direction and strength of their association with CCLE mean viability (from strongest negative to strongest positive). Shaded regions indicate features with significant univariate association (FDR-adjusted $P < 0.10$). Blue indicates negative association with response and red indicates a positive association. The univariate ranks of features selected by the elastic net using gCSI drug response are indicated with vertical bars, also colored by direction of association. In general, the gCSI-selected features are strongly enriched at the extreme ends of the distributions, indicating that they are highly relevant to both drug response data sets. **c**, As for panel **b**, but applying the elastic net to CCLE drug response data and comparing results to univariate statistics obtained from gCSI drug response.

© 2016 Macmillan Publishers Limited. All rights reserved

a **Features identified by elastic net for gCSI and GDSC discretized mean viability data (gCSI genomics)**

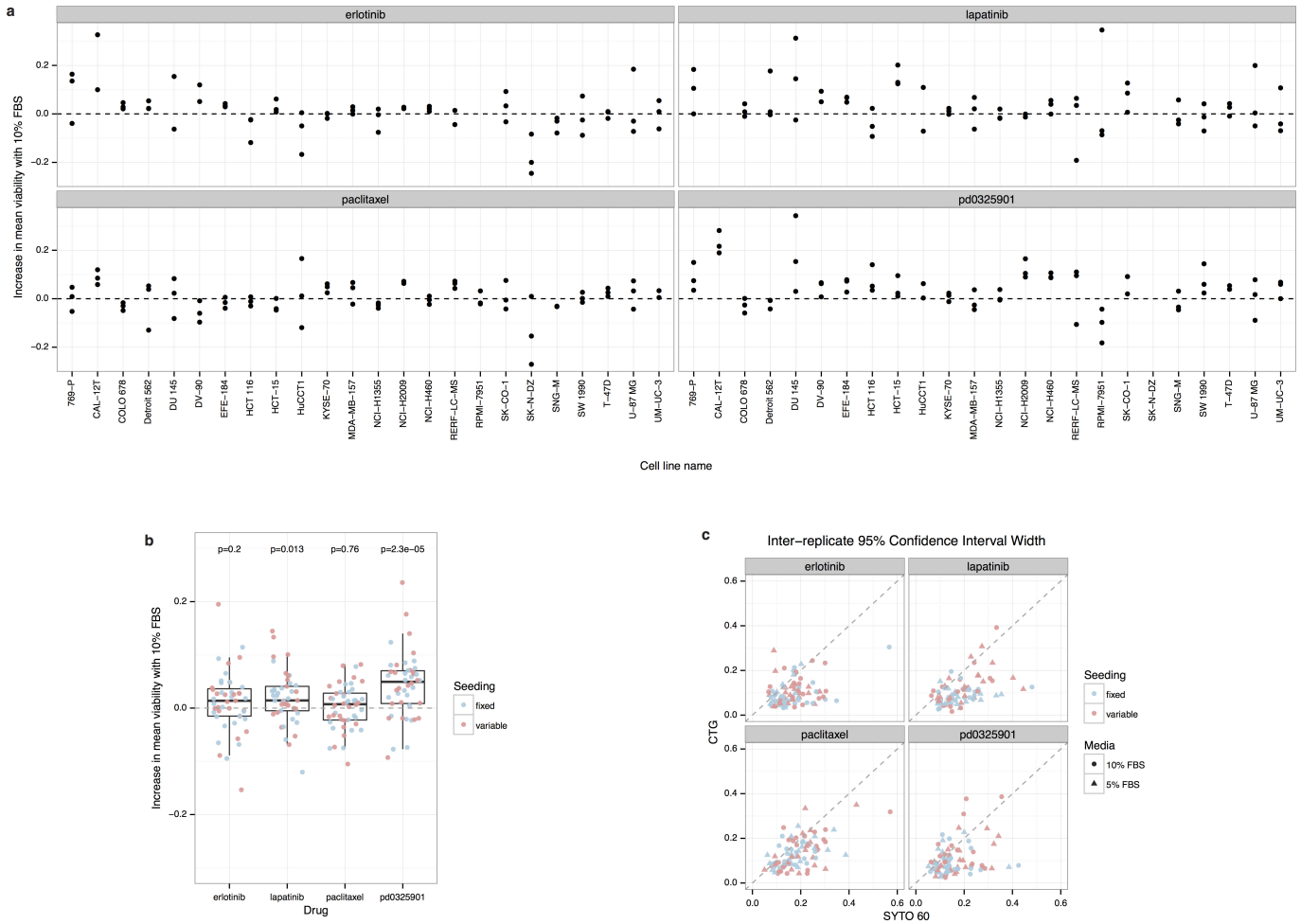
Drug	Target	Expected Biomarkers	Selected positive control features		Total features		Shared Genes
			gCSI	GDSC	gCSI	GDSC	
crizotinib	ALK, MET	MET, HGF expression, MET copy number			3	0	
erlotinib	EGFR	EGFR mutation, expression, or copy number			0	0	
lapatinib	ERBB2	EGFR or ERBB2 copy number or expression, EGFR mutation			1	0	
paclitaxel	Tubulin	ABCB1 expression	ABCB1	ABCB1	8	45	ABCB1
pd0325901	MEK	SPRY, DUSP, ETV4 expression, RAS/RAF mutation	MAPK metafeature, SPRY2	MAPK metafeature	67	106	C16orf71, EBF2, LOC285593, MAPK metafeature, MTCP1NB, NPAS2, PHLDA1, SH3TC2, UCP2





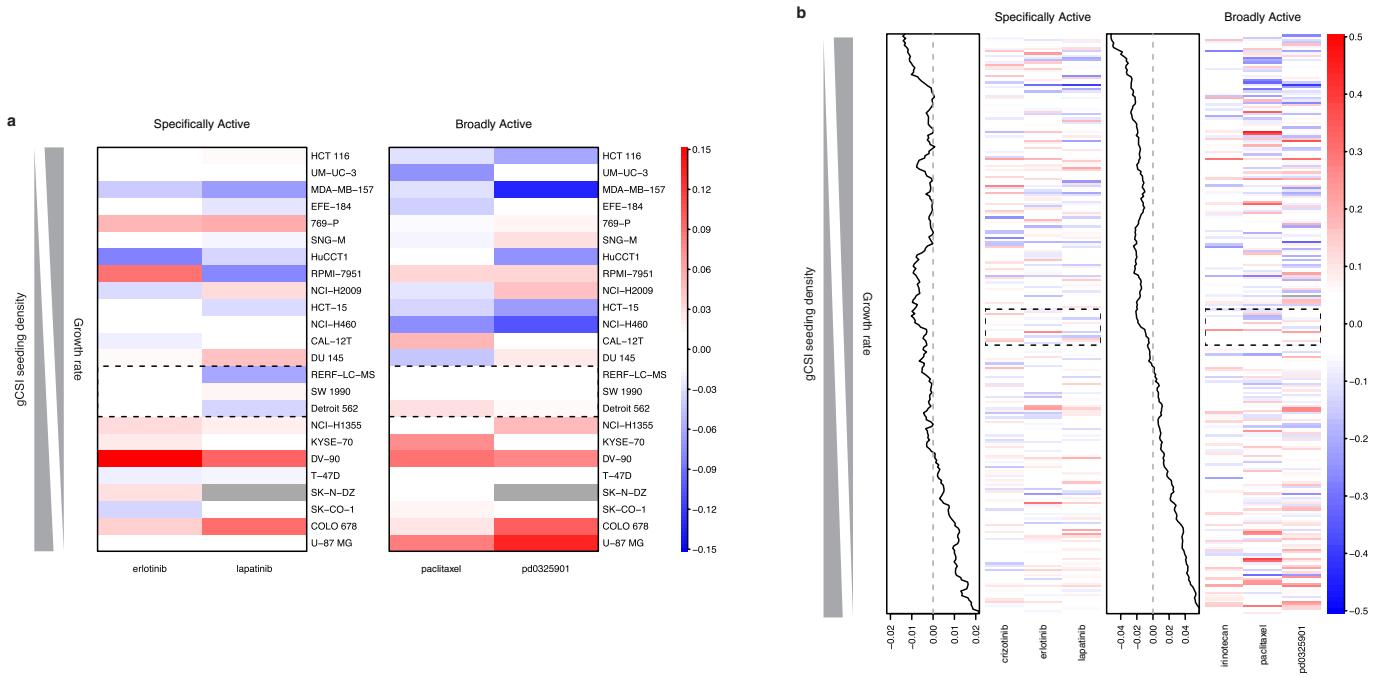
Extended Data Figure 7 | Features identified as response biomarkers using gCSI and GDSC discretized mean viability drug response data versus gCSI genomics data. Equivalent to Extended Data Fig. 6, but contrasting gCSI with GDSC rather than CCLE. **a**, Description of biomarkers expected to associate with response for each of the six drugs tested by both gCSI and GDSC; recovery of expected biomarkers; total selected feature count; and overlap between gCSI and GDSC selected feature sets. **b**, Genomic features were ranked by direction and strength of their association with GDSC mean viability (from strongest negative to strongest positive). Shaded regions indicate features with significant univariate association (FDR-adjusted $P < 0.10$). Blue indicates negative association with response and red indicates a positive association. The univariate ranks of features selected by the elastic net using gCSI drug response are indicated with vertical bars, also colored by direction of association. In general, the gCSI-selected features are strongly enriched at the extreme ends of the distributions, indicating that they are highly relevant to both drug response data sets. **c**, As for panel **b**, but applying the elastic net to GDSC drug response data and comparing results to univariate statistics obtained from gCSI drug response.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 8 | Media serum concentration affects mean viability for some drugs and cell lines. Differences in mean viability between the 10% and 5% FBS conditions are shown for the follow-up screen. Only CTG-derived mean viability values were used. **a**, Changes in single-run mean viability for variable seeding data before synthesizing replicates. In many cases, direction of change was not consistent. In others (for example, 769P or CAL-12T when treated with PD0325901), richer media had a consistent effect in each replicate. For each drug, overall significance of bidirectional effect was assessed by running a paired *t*-test for each cell line separately, and then pooling results by Fisher's method. Overall media effect was significant for all drugs but lapatinib (for erlotinib, $P = 0.037$; lapatinib, $P = 0.288$; paclitaxel, $P = 0.014$; and PD0325901, $P = 2.8 \times 10^{-4}$). **b**, Only PD09325901 achieved an average media effect across cell lines that was statistically significant and of practically important size (4 percentage points). Data for fixed and variable seeding are plotted together because no significant difference in mean viability shift was observed between them for any of the four drugs considered (two-sided, unpaired *t*-tests: erlotinib, $n = 47$, $P = 0.86$; lapatinib, $n = 45$, $P = 0.29$; paclitaxel, $n = 48$, $P = 0.80$; and PD0325901, $n = 46$, $P = 0.44$). Mean viability values were derived from biological triplicate experiments (Methods). **c**, SYTO 60 produces more variable viability measurements than CTG. The width of the 95% confidence intervals for mean viability the 24 cell lines in the follow-on screen (on viability scale; see Methods for confidence interval estimation). Symbol colour indicates seeding strategy; symbol shape indicates media condition. For all drugs and screening conditions, the vast majority of cell lines showed wider confidence intervals for SYTO 60 than for CTG, indicating increased noise in the SYTO 60 data and reduced precision in estimation of mean viability (mean increase for erlotinib was 7 percentage points, $P = 1.2 \times 10^{-11}$; for lapatinib, 8 points, $P = 1.6 \times 10^{-14}$; for paclitaxel, 5 points, $P = 2.7 \times 10^{-10}$; and for PD0325901, 5 points, $P = 2.5 \times 10^{-7}$). This effect is not related to seeding strategy or media serum concentration.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 9 | Seeding density affects drug response for multiple drugs and cell lines. a, In the follow-up screen, the difference in estimated CTG mean viability between variable and fixed seeding is shown (with the two media conditions averaged). For broadly active compounds, slow growing lines typically yielded higher mean viability estimates under variable seeding and therefore produced a positive difference. Fast growing lines, on the other hand, tended to yield lower mean viability estimates under variable seeding (Spearman's ρ test, $P = 1.71 \times 10^{-5}$). This trend was not significant for specifically active drugs ($P = 8.77 \times 10^{-2}$). Lines of average growth rate, for which fixed and variable seeding were essentially the same and thus no difference was expected, are indicated in the dashed rectangle. **b**, Similar to panel **a**, but showing the difference between gCSI (variable seeding) and CCLE (fixed seeding). Because mean viability statistics are not directly comparable when different dosing schedules are used, residuals from a regression of gCSI mean viability on CCLE mean viability are plotted in the heat maps rather than a simple difference. The line plot shows a moving average (75 cell line window) of the row averages of each heat map. As for the follow-up screen, a significant association was seen between cell line growth rate and impact of seeding density for broadly active drugs (Spearman's ρ test, $P = 7.76 \times 10^{-9}$), but not for specifically active drugs ($P = 1.12 \times 10^{-1}$). Mean viability values were derived from biological triplicate experiments (Methods).

© 2016 Macmillan Publishers Limited. All rights reserved

Extended Data Table 1 | Key differences in assay conditions between projects

a **Differences in assay conditions**

Condition	gCSI	CCLE	GDSC
Viability readout	CellTiter-Glo	CellTiter-Glo	SYTO 60
Cell seeding	Variable	Fixed	Variable
Media	5% FBS	10% FBS	5% FBS
Plate format	384	1,536	96 or 384

b **Cell lines available for mean viability comparisons**

Type	Drug	gCSI vs. CCLE	gCSI vs. GDSC
specific	**bid1870**	0	252
broad	**bortezomib**	0	92
specific	**crizotinib**	285	102
broad	**docetaxel**	0	248
broad	**doxorubicin**	0	288
specific	**erlotinib**	281	83
specific	**gdc0941**	0	238
broad	**gemcitabine**	0	286
broad	**irinotecan**	166	0
specific	**lapatinib**	280	96
specific	**ms275**	0	91
broad	**paclitaxel**	271	100
broad	**pd0325901**	281	250
specific	**rapamycin**	0	96
broad	**thapsigargin**	0	289
broad	**vorinostat**	0	242

a, For viability, gCSI and CCLE used CellTiter-Glo, which measures metabolism by quantifying ATP levels; GDSC used the SYTO 60 nucleic acid stain, producing a signal roughly proportional to cell count. GDSC and gCSI varied the initial number of cells plated for each line in order to achieve similar confluency in control wells at the end of the assay; CCLE used a fixed number of cells for every line. GDSC and gCSI supplemented cell culture media with 5% FBS, while CCLE used 10% FBS. All factors except plate format were evaluated in our follow-up screen. Values in red distinguish one data set from the other two. **b**, Cell lines available for mean viability comparisons. The number of available GDSC mean viability or re-calculated CCLE mean viability values were tabulated for all drugs and cell lines that had matching gCSI mean viability data (see Extended Data Fig. 1).

© 2016 Macmillan Publishers Limited. All rights reserved

ARTICLE

doi:10.1038/nature17967

A platform for the discovery of new macrolide antibiotics

Ian B. Seiple[1]*†, Ziyang Zhang[1]*, Pavol Jakubec[1], Audrey Langlois-Mercier[1]†, Peter M. Wright[1]†, Daniel T. Hog[1]†, Kazuo Yabu[1]†, Senkara Rao Allu[1], Takehiro Fukuzaki[1], Peter N. Carlsen[1], Yoshiaki Kitamura[1]†, Xiang Zhou[1], Matthew L. Condakes[1]†, Filip T. Szczypiński[1]†, William D. Green[1]† & Andrew G. Myers[1]

The chemical modification of structurally complex fermentation products, a process known as semisynthesis, has been an important tool in the discovery and manufacture of antibiotics for the treatment of various infectious diseases. However, many of the therapeutics obtained in this way are no longer effective, because bacterial resistance to these compounds has developed. Here we present a practical, fully synthetic route to macrolide antibiotics by the convergent assembly of simple chemical building blocks, enabling the synthesis of diverse structures not accessible by traditional semisynthetic approaches. More than 300 new macrolide antibiotic candidates, as well as the clinical candidate solithromycin, have been synthesized using our convergent approach. Evaluation of these compounds against a panel of pathogenic bacteria revealed that the majority of these structures had antibiotic activity, some efficacious against strains resistant to macrolides in current use. The chemistry we describe here provides a platform for the discovery of new macrolide antibiotics and may also serve as the basis for their manufacture.

Natural products have provided critical starting points for the development of a majority of the antibacterial drugs listed as essential medicines by the World Health Organization. During the period from about 1940 to 1960, sometimes described as the golden age of antibiotics research, academic and industrial laboratories identified the natural products that went on to define many of the major classes of modern antibiotics[1]. Few if any natural products prove to be optimal for safety, efficacy or oral use in humans, however, for these were probably not evolutionary pressures for the microbes in which they developed. For six decades a primary tool by which new antibiotics have been discovered and manufactured is semisynthesis, or the chemical modification of natural products derived from fermentation. Semisynthesis is inherently limited because it is challenging to modify structurally complex natural products selectively, and typically few positions of any given scaffold can be modified effectively[2,3]. Macrolide antibiotics (macrocyclic lactones with one or more pendant glycosidic residues), which have proven to be safe and effective for use in treating human infectious diseases such as community-acquired bacterial pneumonia, gonorrhoea and others, provide a compelling case in point. Since the discovery of erythromycin from a Philippine soil sample in 1949[4], in spite of extraordinary efforts leading to fully synthetic[5–8] and modified biosynthetic routes[9–12] to macrolide antibiotics, all members of this class approved or in clinical development for use in humans have been manufactured by chemically modifying erythromycin (erythromycin itself is unstable in the stomach, and rearranges to form a product with gastrointestinal side-effects)[13]. Thus, azithromycin is prepared from erythromycin in four steps[14], clarithromycin requires six steps[15], and the advanced clinical candidate solithromycin is currently produced from erythromycin by a 16-step linear sequence (Fig. 1)[16,17]. New macrolides are badly needed, for resistance to approved macrolide antibiotics such as clarithromycin and azithromycin is now widespread in both hospitals and the community[18–20]. Here we present a practical, fully synthetic platform for the preparation of macrolide antibiotics, providing both a discovery engine for structurally novel antibiotic candidates that would be difficult or impossible to obtain from erythromycin, as well as a basis for their eventual manufacture. Using simple building blocks and a highly convergent assembly process, we have prepared more than 300 structurally diverse macrolide antibiotic candidates, as well as the approved drug telithromycin and the clinical candidate solithromycin. We have identified molecules with diverse macrocyclic scaffolds that exhibit potent activities against bacterial strains resistant to erythromycin, azithromycin and other current antibiotics of different classes.

Synthesis of 14-membered azaketolides

To illustrate our approach to the synthesis of macrolide antibiotics, we begin by detailing a route to new and highly active hybrid antibiotics, 14-membered azaketolides (Fig. 2), and then show how the approach can be easily expanded to subsume many other active macrolide scaffolds, with broad latitude for substitutional variation. Ketolides replace the C3-cladinose carbohydrate residue of erythromycin with a carbonyl group, a chemical modification that allows them to evade certain inducible forms of macrolide resistance[21,22], whereas the azalide azithromycin incorporates a nitrogen atom within an expanded (15-membered) macrolactone ring, features that have been proposed to impart favourable pharmacological properties, including increased efficacy against certain Gram-negative pathogens[23,24]. The hybrid 'azaketolides', or macrolides that contain both a C3 carbonyl group and a nitrogen atom within the macrolactone ring, have been little explored; one such compound has been reported and was found to be essentially inactive[25]. We were particularly interested in 14-membered azaketolides that would arise from formal replacement of the carbonyl group at position C9

[1]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, USA. †Present addresses: Department of Pharmaceutical Chemistry, University of California, San Francisco, California 94158, USA (I.B.S.); Novartis Pharma AG, Chemical and Analytical Development, CH-4002 Basel, Switzerland (A.L.-M.); McKinsey and Company, 55 East 52nd Street, 21st Floor, New York, New York 10022, USA (P.M.W.); Bayer Pharma AG, Medicinal Chemistry, Müllerstrasse 178, 13353 Berlin, Germany (D.T.H.); Medicinal Chemistry Research Laboratories, Daiichi Sankyo Co., Ltd, Shinagawa R&D Center, 1-2-58 Hiromachi, Shinagawa, Tokyo 140-8710, Japan (K.Y.); Department of Chemistry and Biomolecular Science, Gifu University, 1-1 Yanagido, Gifu 501-1193, Japan (Y.K.); Department of Chemistry, University of California, Berkeley, California 94720, USA (M.L.C.); Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK (F.T.S.); Trinity College, University of Cambridge, Cambridge CB2 1TQ, UK (W.D.G.).
*These authors contributed equally to this work.

© 2016 Macmillan Publishers Limited. All rights reserved

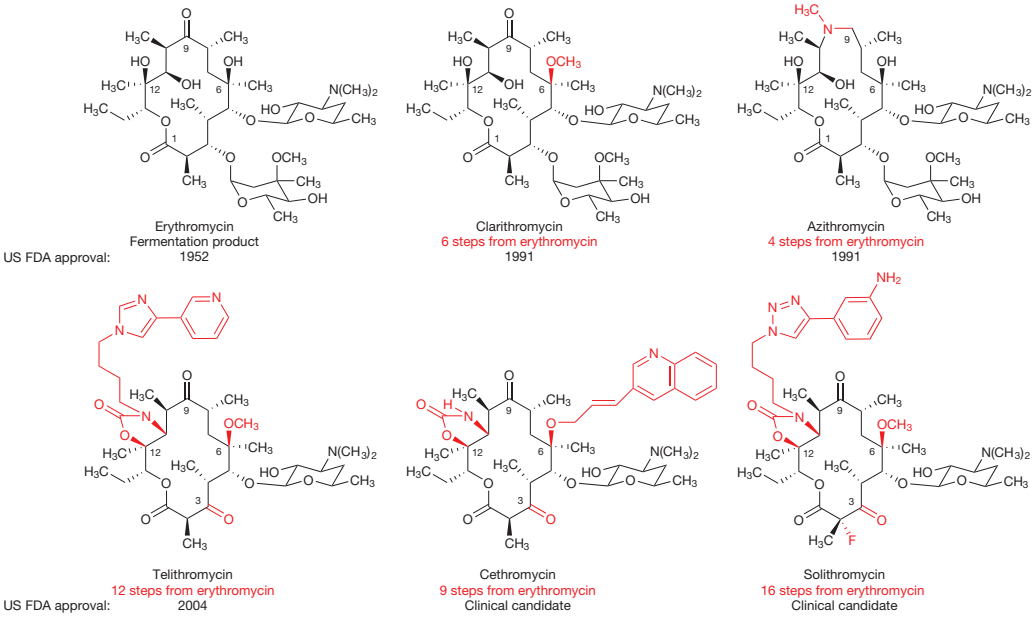


Figure 1 | Summary of macrolide antibiotic development by semisynthesis. To date, all macrolide antibiotics are produced by chemical modification (semisynthesis) of erythromycin, a natural product produced on the ton scale by fermentation. Depicted are erythromycin and the approved semisynthetic macrolide antibiotics clarithromycin, azithromycin and telithromycin along with the dates of their FDA approval and the number of steps for their synthesis from erythromycin. The previous ketolide clinical candidate cethromycin and the current clinical candidate solithromycin are also depicted. It is evident that increasingly lengthy sequences are being employed in macrolide discovery efforts.

of ketolides with an amino group, a transformation that would not be feasible by semisynthesis.

An overview is helpful, as it emphasizes our synthetic strategy. Eight simple building blocks (**1**–**8**), two of which are commercially available and six of which are synthesized in sequences of 3–6 steps in 30–150-g amounts, are assembled in a highly convergent manner (Fig. 2). As depicted, building blocks **1**–**3** are transformed by a seven-step sequence featuring two convergent coupling reactions to provide a key left-hand intermediate (**15**, an amine) in 57% yield, while building blocks **4**–**7** are transformed in a seven-step sequence featuring three convergent coupling reactions to form a key right-hand intermediate (**22**, a ketone) in 58% yield. (Here left and right refer to the position in the graphical representation of the target molecule, Fig. 2a.) The left- and right-hand intermediates are then joined in a highly stereoselective and efficient reductive coupling reaction (82% yield, diastereomeric ratio (d.r.) >20:1) that also serves to remove the benzoate protective group within the desosamine residue. The next step is the key macrocyclization reaction, for which we employ a method developed by Boeckman *et al.* in the context of their approaches towards the natural products diplodialide A (a 10-membered macrolactone), kromycin (a 14-membered macrolactone) and ikarugamycin (a 17-membered macrolactam)[26,27]. We found that thermolysis of intermediate **23** (1 mM in chlorobenzene) at 132 °C afforded the macrolactone **24** in 78% yield (1-g scale). There are a number of notable features of this transformation, not least its efficacy and, as we will show, generality. Many pioneering efforts to synthesize macrolide antibiotics failed at the stage of macrocyclization; successful macrocyclizations were shown to require carefully designed, rigidifying protective group strategies[28]. We believe that the success of the present cyclization (which we note occurs in the presence of a free secondary amine as well as another secondary alcohol) is attributable in part to the rigid structuring induced by the cyclic carbamate function on the left-hand side of the molecule, which we suspect brings the putative acyl ketene function (produced from the dioxolenone function upon thermolysis) and the reactive secondary alcohol into favourable proximity. A single operation then transforms the product of macrocyclization into a fully synthetic antibiotic candidate (**25**) while incorporating the final building block using Sharpless' optimized conditions for azide–alkyne dipolar cycloaddition[29] (**8**, 86% yield). This final step is by design a convergent coupling reaction, and permits optimal late-stage diversification of a key side-chain residue that is known to occupy a novel pocket in the bacterial ribosome[30–34].

Aspects of certain steps before the macrocyclization reaction merit brief discussion (see Fig. 2). The key aldol coupling reaction that unites building blocks **1**[35] and **2**[36], initiating the synthesis of the left-half intermediate, involves new methodology that was developed specifically for the present purpose (though it proves to be a general transformation) and proceeds in 98% yield on a 33-g scale to form a single, crystalline diastereomer[36]. In a subsequent step, which transforms the pseudoephenamine amide function into a methyl ketone in one operation, the pseudoephenamine auxiliary is recovered by a simple aqueous extraction procedure in 92% yield. The convergent coupling reaction that unites building blocks **4** and **5** to begin the construction of the right-half intermediate is based on a palladium-catalysed ene reaction developed by Mikami and co-workers, and proceeds in 93% yield and 92% enantiomeric excess (e.e.)[37]. Two convergent coupling reactions follow three linear functional group interconversions, the first coupling being a Mukaiyama[38] aldol reaction (to incorporate building block **6**[39]) whose diastereoselectivity was observed to be dependent upon the reaction conditions. The conditions employed here (magnesium bromide etherate as Lewis acid, triethylamine as base, dichloromethane as solvent, and a reaction temperature of −78 °C) provided maximum selectivity (d.r. >20:1, 91% yield) for the desired syn-aldol product. The next convergent coupling reaction introduced the protected desosamine carbohydrate residue and for this purpose we employed the thioglycoside **7**[40,41], a donor very similar to that employed by Woodward *et al.* in their synthesis of erythromycin[5]. In the present case, we found that the use of a benzoyl group to shield the secondary alcohol on the glycosidic donor in lieu of methoxycarbonyl improved the yield and diastereoselectivity of the glycosidation reaction. The overall yield of the fully synthetic azaketolide antibiotic candidate **25** was 32% from initial building blocks **1** and **2** (a 10-step sequence) or 32% from initial building blocks **4** and **5** (also a 10-step sequence).

Synthesis of 15-membered azaketolides

By design, the convergent building block strategy we employ enables rapid access to structurally related scaffolds; for example, synthesis of a 15-membered hybrid azaketolide scaffold comprising structural

© 2016 Macmillan Publishers Limited. All rights reserved

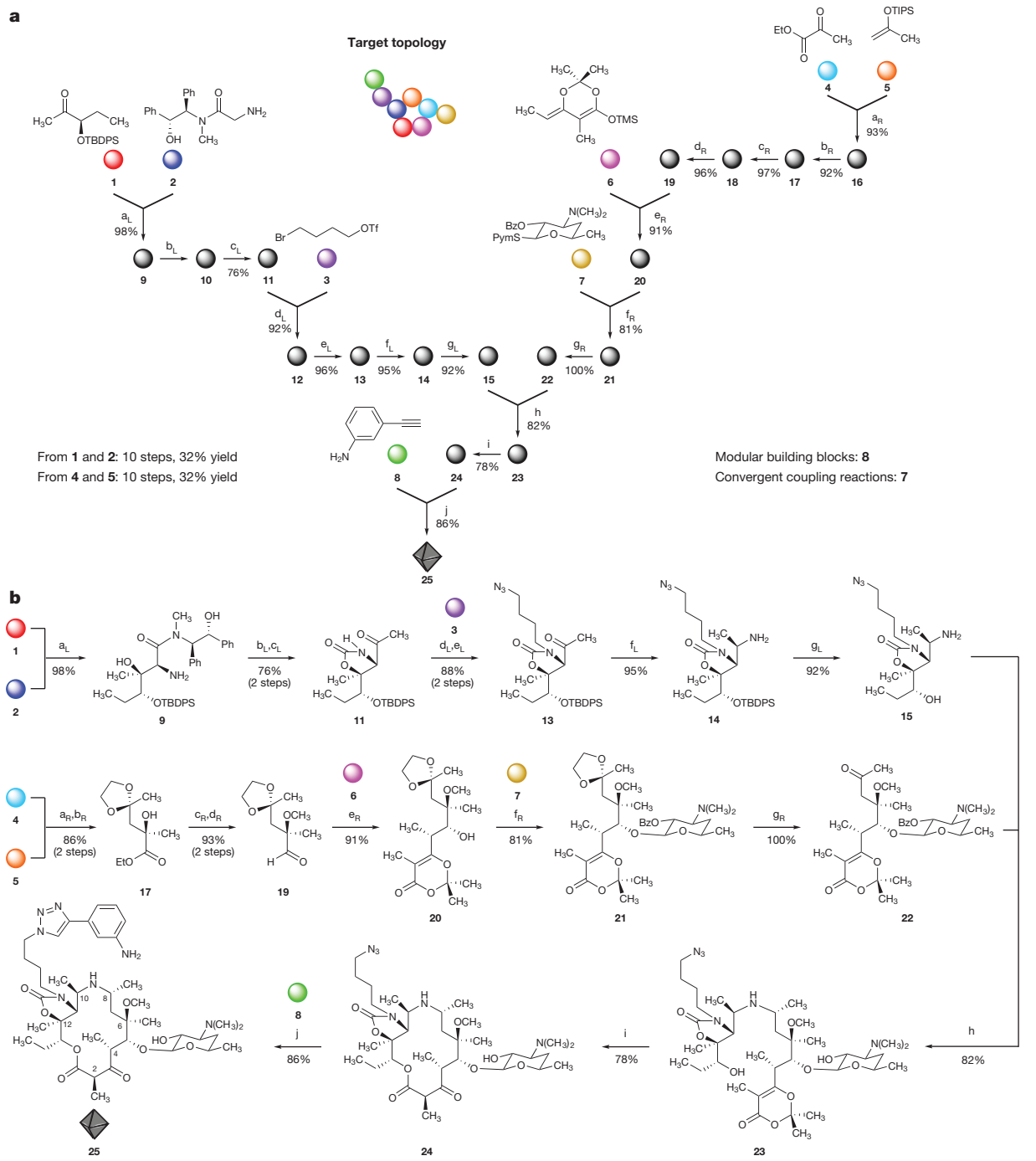


Figure 2 | A convergent, fully synthetic route to the 14-membered azaketolide 25. **a**, Graphical representation of the convergent synthesis of azaketolide **25** from eight variable building blocks (represented by coloured spheres). Downward, 'Y'-shaped arrows signify convergent coupling reactions. **b**, Synthesis of azaketolide **25**, reagents and conditions (subscripts L and R indicate left and right halves, respectively) as follows: (a_L) LiHMDS, LiCl, 98%; (b_L) Et^{i}Pr$_2$N, $COCl_2$; (c_L) ^{i}PrMgCl, CH_3Li, 76% over 2 steps, 92% of recovered (*R*,*R*)-pseudoephenamine; (d_L) NaHMDS; (e_L) NaN_3, 88% over 2 steps; (f_L) NH_3, Ti(O^{i}Pr)$_4$, $NaBH_4$, 95%; (g_L) Bu_4NF, 92%; (a_R) Pd[(*S*)-SegPhos]Cl_2, $AgSbF_6$, 93%, 92% e.e.; (b_R) ethylene glycol, PPTS, 92%; (c_R) KH, MeI, 97%; (d_R) (^{i}Bu)$_2$AlH, 96%; (e_R) Et_3N, $MgBr_2$•OEt_2, 91%; (f_R) AgOTf, 81%, 16:1 β:α; (g_R) HCl, 100%; (h) $NaCNBH_3$, 82%; (i) 132 °C, 1 mM, PhCl, 78%; (j) $CuSO_4$, sodium L-ascorbate, 86%.

features of both the azalide antibiotic azithromycin and the ketolide clinical candidate solithromycin (Fig. 3) required modification of just three of the eight building blocks that were used to assemble 14-membered azaketolides (Fig. 2). The left-hand amine intermediate (**15**) was the same in both 14- and 15-membered azaketolide scaffold syntheses and thus construction of the 15-membered azaketolides simply required access to the homologated right-hand intermediate, aldehyde **35**. Here, too, we employed a seven-step sequence (38% yield) featuring three convergent coupling reactions to assemble four simple building blocks. The first coupling, which unites building blocks **26**[42] and **27**[43,44] to form the tertiary alcohol **29**, emerged only after extensive experimentation, in which we discovered that incorporating the large *tert*-butyldiphenylsilyl protective group within component **27** was key to achieve high diastereoselectivity. After two subsequent functional group interconversions, building blocks **6** (as before) and **28** (the original glycosyl donor of Woodward *et al.*[5], which proved to be effective in this scaffold synthesis) were incorporated by successive convergent coupling reactions. Two linear steps then provided the right-hand

© 2016 Macmillan Publishers Limited. All rights reserved

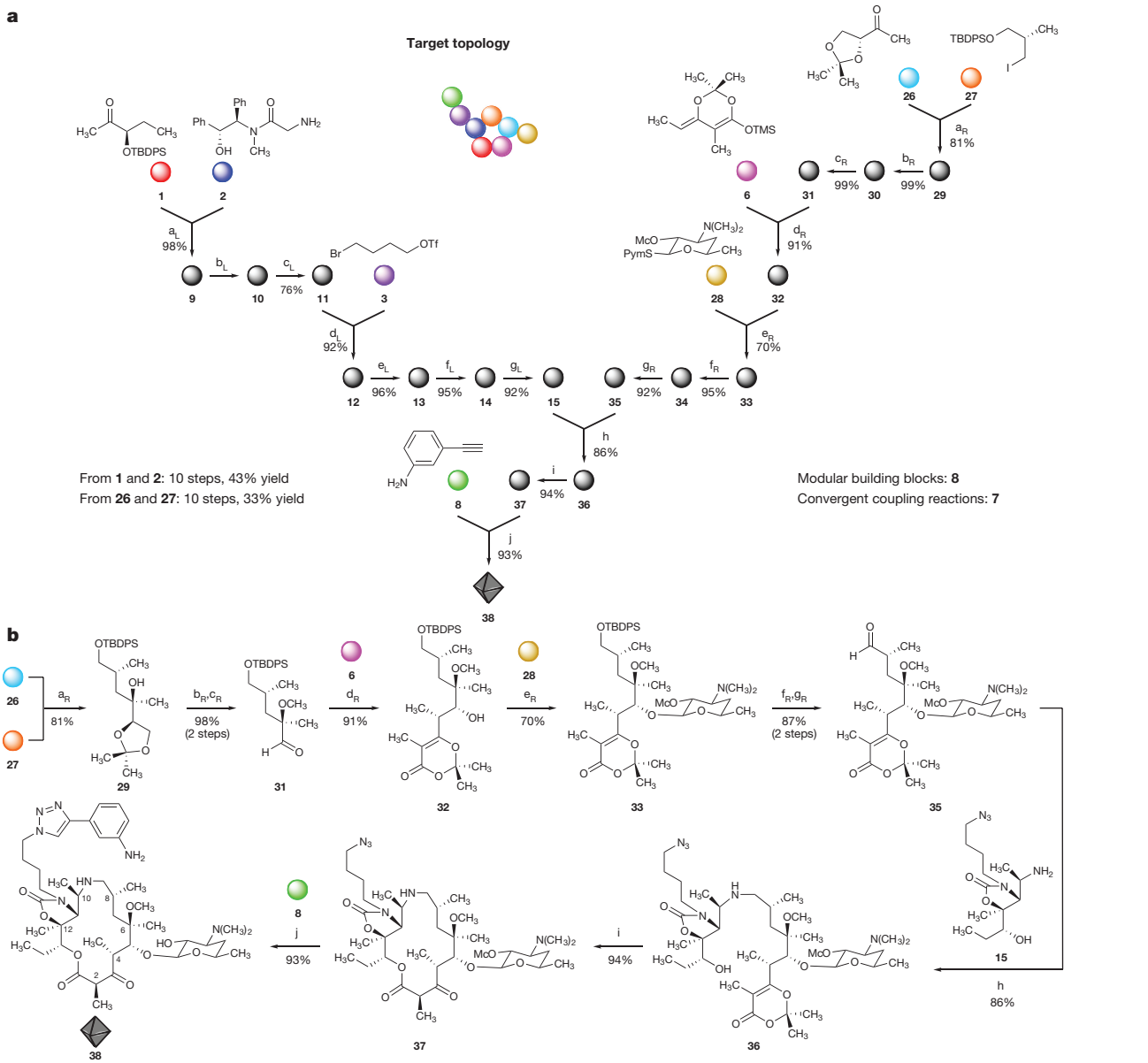


Figure 3 | A convergent, fully synthetic route to the 15-membered azaketolide 38. a, Graphical representation of the convergent synthesis of azaketolide **38** from eight variable building blocks (represented by coloured spheres). Downward, 'Y'-shaped arrows signify convergent coupling reactions. **b**, Synthesis of azaketolide **38**, reagents and conditions (subscripts L and R indicate left and right halves, respectively): (a_L) LiHMDS, LiCl, 98%; (b_L) Et^{i}Pr$_2$N, $COCl_2$; (c_L) ^{i}PrMgCl, CH_3Li, 76% over 2 steps, 92% of recovered (*R*,*R*)-pseudoephenamine; (d_L) NaHMDS; (e_L) NaN_3, 88% over 2 steps; (f_L) NH_3, Ti(O^{i}Pr)$_4$, $NaBH_4$, 95%; (g_L) Bu_4NF, 92%; (a_R) ^{t}BuLi, $MgBr_2$, 81%; (b_R) KH, MeI, 99%; (c_R) H_5IO_6, 99%; (d_R) $ZnCl_2$, 91%; (e_R) AgOTf, 70%; (f_R) HF (aq.); (g_R) Dess–Martin periodinane, 87% over 2 steps; (h) $NaCNBH_3$, 86%; (i) 132 °C, 1 mM, PhCl, 94%; (j) $CuSO_4$, sodium L-ascorbate, 93%.

intermediate, aldehyde **35**. Reductive coupling of the left-hand amine (**15**) and the right-hand aldehyde (**35**) in the presence of sodium cyanoborohydride proceeded without detectable epimerization to afford the macrocyclization precursor **36** in 86% yield. Thermolysis of **36** at 132 °C in chlorobenzene then afforded the 15-membered macrolide **37** in 94% yield on a 1.9-g scale. Incorporation of the final building block (**8**, as before) and cleavage of the methoxycarbonyl protective group were achieved in a single, final step to provide the azaketolide antibiotic candidate **38**. The overall yield of the fully synthetic macrolide **38** (FSM-20707, see below) was 43% from initial building blocks **1** and **2** (a 10-step linear sequence) or 33% from initial building blocks **26** and **27** (also a 10-step linear sequence).

Synthesis of 14-membered ketolides

We next adapted our convergent assembly strategy to synthesize 14-membered ketolides, a class that includes the clinical candidate solithromycin as well as the approved drug telithromycin, with the goal of preparing diverse structural analogues. Towards this end we envisioned coupling of the same right-hand intermediate employed in the synthesis of 15-membered azaketolides (aldehyde **35**) with a Grignard reagent derived from hydromagnesiation of the acetylenic alcohol **42** (Fig. 4). Intermediate **42** was obtained in one step (14:1 d.r., 85% yield) by the addition of 1-lithiopropyne (**39**) to ketone **40**[35] in the presence of lithium (1*S*,2*R*)-1-phenyl-2-(pyrrolidin-1-yl)-1-propanolate[45]. Hydromagnesiation of **42** with cyclopentylmagnesium bromide and bis-cyclopentadienyltitanium dichloride[46] proceeded regioselectively to furnish a vinyl Grignard reagent that was captured *in situ* with aldehyde **35**. The resulting diastereomeric mixture of allylic alcohols (obtained in 80% yield on a 3-g scale) was oxidized directly (Dess–Martin periodinane, 97%) and the α,β-unsaturated ketone produced was desilylated (tetrabutylammonium fluoride (Bu_4NF), 95%) to afford the keto alcohol **45**. Macrocyclization, as before, then provided the 14-membered

© 2016 Macmillan Publishers Limited. All rights reserved

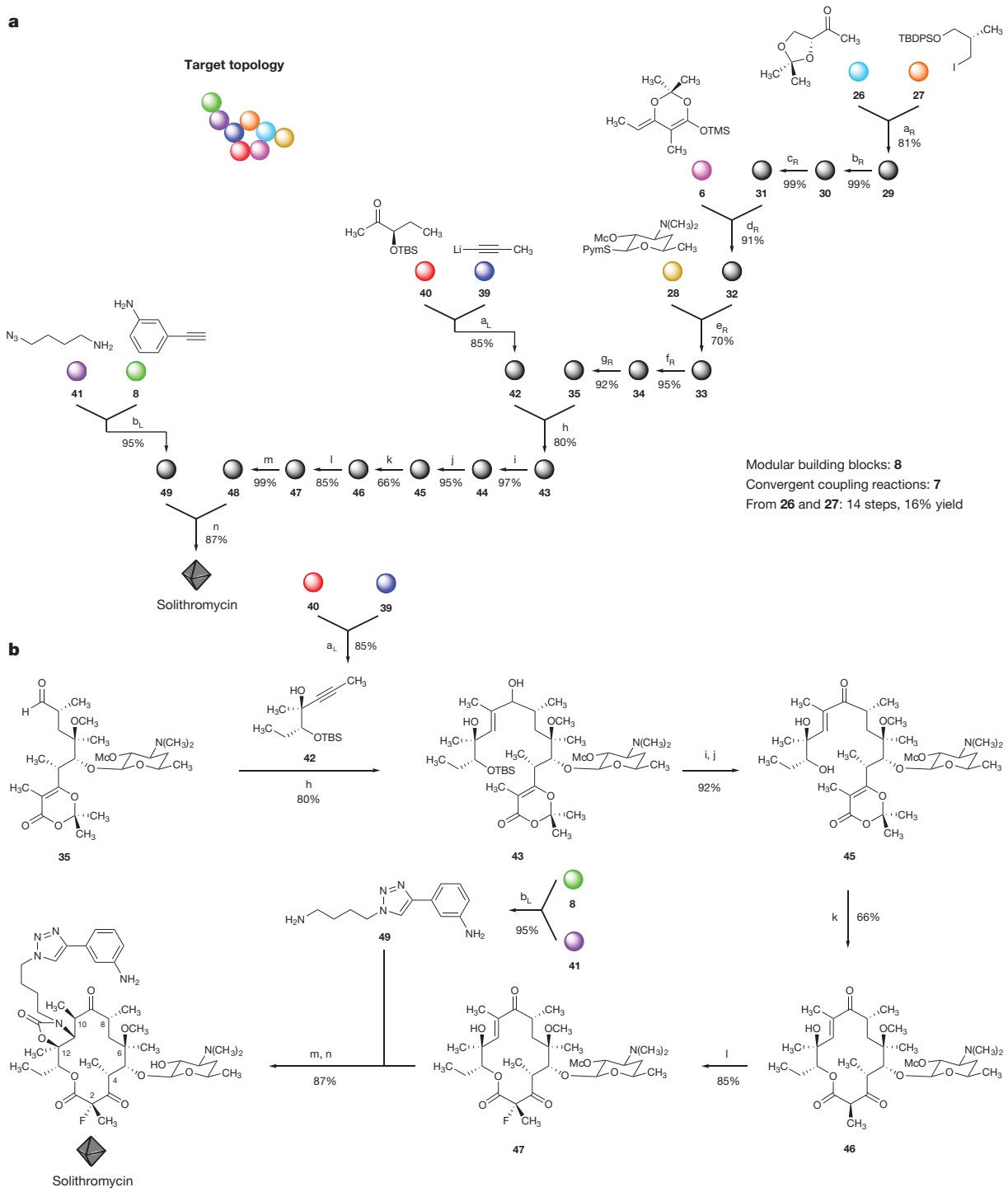
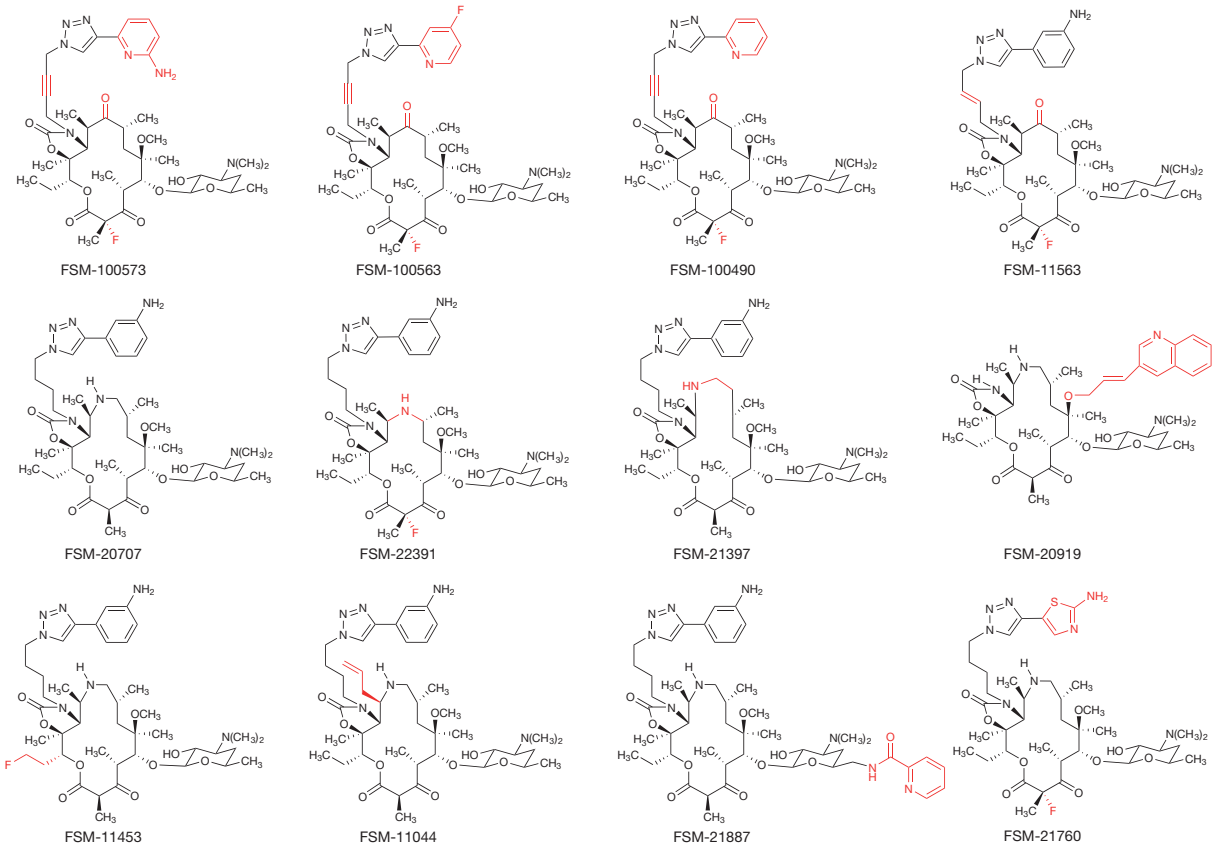


Figure 4 | A convergent, fully synthetic route to solithromycin. **a**, Graphical representation of the convergent synthesis of solithromycin, which was previously only accessible by semisynthesis. This route has been adapted for the synthesis of >30 novel ketolide antibiotic candidates, as well as the FDA-approved ketolide telithromycin. Downward, 'Y'-shaped arrows signify convergent coupling reactions. **b**, Synthesis of solithromycin, reagents and conditions (subscripts L and R indicate left and right portions, respectively): (a_L) lithium (1*S*,2*R*)-1-phenyl-2-(pyrrolidin-1-yl)-1-propanolate, 85%; (b_L) $CuSO_4$, sodium L-ascorbate, 95%; (a_R) ^{t}BuLi, $MgBr_2$, 81%; (b_R) KH, MeI, 99%; (c_R) H_5IO_6, 99%; (d_R) $ZnCl_2$, 91%; (e_R) AgOTf, 70%; (f_R) HF (aq.), 95%; (g_R) Dess–Martin periodinane, 92%; (h) Cp_2TiCl_2, cyclopentylmagnesium bromide, 80%; (i) Dess–Martin periodinane, 97%; (j) Bu_4NF, 95%; (k) 132 °C, 0.5 mM, PhCl, 66%; (l) KO^{t}Bu, $FN(SO_2Ph)_2$, 85%; (m) Im_2CO, DBU; (n) imidazole hydrochloride, 60 °C, 87% over 2 steps.

macrolactone **46** in 66% yield (1.7-g scale). Fluorination at position C2 proceeded in 85% yield[47]. Transformation of the C12 tertiary alcohol within the latter intermediate (**47**) to the corresponding acyl imidazolide (carbonyl diimidazole (Im_2CO), 1,8-diazabicyclo[5.4.0]undec-7-ene (DBU)) and trapping with amine **49** then afforded fully synthetic solithromycin in one operation (87% yield, the methoxycarbonyl protective group was cleaved concomitantly under these conditions).

The present route to solithromycin proceeds in 14 steps and 16% yield from building blocks **26** and **27**, and differs from the semisynthetic route (16 steps from erythromycin, yields not reported[16]) in that the linker and its attached side-chain heterocycle are introduced in the final step of our synthesis and thus modification of these residues is quite facile, whereas the linker residue is introduced 6 steps before the penultimate step in the published semisynthetic route to solithromycin[16].

© 2016 Macmillan Publishers Limited. All rights reserved



	Species	Strain description	Erythro	Azithro	Telithro	Solithro	100573	100563	100490	11563	20707	22391	21397	20919	11453	11044	21887	21760
Gram-positive	*S. aureus*	ATCC 29213	0.5	1	0.125	0.125	0.06	≤0.03	≤0.03	0.06	0.5	0.25	4	0.25	1	1	8	0.5
	S. aureus	BAA-977; iErmA	>256	>256	0.06	≤0.03	0.06	0.06	0.03	0.06	0.5	0.5	4	0.5	1	1	8	1
	S. aureus	MP513; MRSA; cErmA	>256	>256	256	>64	16	16	64	64	>64	64	64	64	>64	>64	>64	64
	S. aureus	NRS384; MRSA; MsrA	64	128	0.125	0.25	0.06	0.125	0.06	0.125	1	1	8	0.5	2	2	16	4
	S. pneumoniae	ATCC 49619	0.03	0.06	≤0.03	≤0.03	≤0.03	≤0.03	≤0.03	≤0.03	≤0.03	≤0.03	0.06	≤0.03	≤0.03	≤0.03	0.06	≤0.03
	S. pneumoniae	UNT-042; ErmB/MefA	>256	>256	0.125	0.25	≤0.03	≤0.03	≤0.03	≤0.03	2	0.125	8	0.5	2	8	1	1
	S. pyogenes	ATCC 19615	≤0.03	≤0.03	≤0.03	≤0.03	≤0.03	≤0.03	≤0.03	≤0.03	≤0.03	≤0.03	0.06	≤0.03	≤0.03	≤0.03	0.06	≤0.03
	E. faecalis	ATCC 29212	1	4	≤0.03	≤0.03	0.03	0.03	0.03	≤0.03	0.125	0.06	0.5	0.25	0.125	0.125	0.5	0.06
	E. faecalis	UNT-047; VRE; ErmB	>256	>256	16	32	1	2	2	4	>64	32	64	>64	>64	64	>64	>64
Gram-negative	*H. influenzae*	ATCC 49247	4	2	2	4	2	2	2	2	2	4	8	4	4	8	16	4
	A. baumannii	ATCC 19606	16	32	4	16	2	8	8	4	4	4	16	16	4	32	32	32
	K. pneumoniae	ATCC 10031	4	2	4	4	2	8	4	4	2	4	8	16	2	8	8	4
	E. coli	ATCC 25922	64	4	16	32	8	16	16	16	4	8	32	4	8	64	16	8
	P. aeruginosa	ATCC 27853	64	64	64	64	16	32	64	32	64	64	64	64	>64	>64	>64	64

MIC colour scale (μg ml^{-1}): <0.03 | 0.03 | 0.06 | 0.125 | 0.25 | 0.5 | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | >256

Figure 5 | Structures of selected fully synthetic macrolide antibiotics, and their MICs compared to those of natural and semisynthetic macrolide antibiotics. Top three rows, structural formulae. Bottom row, table showing MICs (values in μg ml^{-1} colour coded, key at bottom) of current macrolide antibiotics and of selected analogues against 9 Gram-positive and 5 Gram-negative microorganisms. iErmA, inducible erythromycin ribosome methyltransferase A; cErmA, constitutive erythromycin ribosome methyltransferase A; MsrA, macrolide streptogramin resistance efflux pump A; MefA, macrolide efflux protein A; ErmB, erythromycin ribosome methyltransferase B; Erythro, erythromycin; Azithro, azithromycin; Telithro, telithromycin; Solithro, solithromycin.

We prepared a number of fully synthetic 14-membered ketolides by taking advantage of this feature of our synthesis to simultaneously vary the linker and side-chain heterocycle, and many of these products show potencies that are superior to solithromycin in microbiological assays (see below).

Construction of a library of macrolide antibiotics

We prepared an initial library of >300 fully synthetic macrolide (FSM) antibiotic candidates by varying the building blocks in concert with modifying readily diversifiable elements (for example, an azido group, an amino group, a β-keto lactone function, an allyl group) that we introduced into and around the macrolide ring, a powerful tactical combination. In addition to members of the three primary macrocyclic scaffolds discussed in detail above, we prepared a number of other unique scaffolds by modifying the principal coupling components (left- and right-halves) and their modes of coupling using straightforward alternative chemical transformations (see Supplementary Information for a complete list of structures synthesized and Extended Data Figs 1–10 for exemplary schemes for their preparation). This strategy not only provided novel scaffolds to explore, but also permitted deep-seated variations of positions within these scaffolds, thus enabling access to molecules that could not be prepared using semisynthetic methods.

Microbiological testing

To evaluate our FSMs for antibiotic activity, we screened 305 of them against a panel of pathogens comprising standardized Gram-positive and Gram-negative strains (see Supplementary Information). For the most promising antibiotic candidates, the panel was expanded to include bacteria with genetically characterized resistance mechanisms to erythromycin as well as other antibiotics. These include three different strains of *Staphylococcus aureus*, two of them clinically isolated methicillin-resistant (methicillin-resistant *Staphylococcus aureus*, MRSA) pathogens, a clinically isolated strain of *Streptococcus*

© 2016 Macmillan Publishers Limited. All rights reserved

pneumoniae with both ribosome-modifying methyltransferase (*ermB*) and efflux (*mefA*) genes, and finally a clinically isolated strain of *Enterococcus faecalis* with an *ermB* gene, a strain resistant to all approved macrolides as well as vancomycin (vancomycin-resistant *Enterococcus*, VRE) (Fig. 5). The data show that a majority of compounds in the library exhibit demonstrable antibiotic activity. For example, 83% of compounds in the collection displayed a minimum inhibitory concentration (MIC) $\leq 4\,\mu g\,ml^{-1}$ against wild-type *S. pneumoniae*, which is known to be highly susceptible to inhibition by macrolides (see Supplementary Information). Among the most promising compounds (Fig. 5) are represented variously 14-, 15- and 16-membered azaketolide scaffolds (for example, FSM-22391, FSM-20707 and FSM-21397, respectively), a 15-membered azacethromycin hybrid antibiotic (FSM-20919), and a number of fully synthetic 14-membered ketolides (most notably, FSM-100573 and FSM-100563). The last two substances display superior potencies relative to any macrolide in current clinical use in the following extremely challenging strains: an *S. pneumoniae* with both *ermB* and *mefA* genes (MIC for both FSM-100573 and FSM-100563 $\leq 0.03\,\mu g\,ml^{-1}$), a VRE with an *ermB* gene (MICs 1 and $2\,\mu g\,ml^{-1}$, respectively), an MRSA with a constitutively-expressed erythromycin ribosome methyltransferase (c-*erm*) gene (MICs $16\,\mu g\,ml^{-1}$), and *Pseudomonas aeruginosa* (MICs 16 and $32\,\mu g\,ml^{-1}$, respectively). Clearly, further work will be required to drive potencies in the latter two strains to clinically efficacious levels, but we believe that demonstration of even modest activity in these challenging strains by a macrolide is significant. It is interesting and encouraging to note that FSM-100573 also displays improved Gram-negative activity relative to many other macrolides; further advances in this regard would address an important area of unmet medical need[48].

Conclusion

Employing as a design strategy the multiply convergent assembly of simple chemical building blocks, we have developed a platform of unprecedented versatility for the discovery and practical synthesis of novel macrolide antibiotics. Varying the building blocks as well as (at a later stage) diversifiable elements incorporated within them permits an almost exponential expansion of variability within any given scaffold. In addition, our work shows that variant scaffolds can be obtained by straightforward perturbations of our generalized assembly process. As with our earlier convergent synthesis of tetracycline antibiotics[49,50], we anticipate that many thousands of novel macrolide structures can be prepared for evaluation as potential antibiotics using the present synthetic platform. In light of this, it seems logical to conclude that developing similar convergent routes to other naturally occurring antibiotic families may accelerate the discovery of new therapeutic agents for human infectious diseases.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 7 December 2015; accepted 23 March 2016.

1. Walsh, C. *Antibiotics: Actions, Origins, Resistance* (American Society for Microbiology Press, 2003).
2. Fischbach, M. A. & Walsh, C. T. Antibiotics for emerging pathogens. *Science* **325,** 1089–1093 (2009).
3. Wright, P. M., Seiple, I. B. & Myers, A. G. The evolving role of chemical synthesis in antibacterial drug discovery. *Angew. Chem. Int. Ed.* **53,** 8840–8869 (2014).
4. Mcguire, J. M. *et al.* Ilotycin, a new antibiotic. *Antibiot. Chemother.* **2,** 281–283 (1952).
5. Woodward, R. B. *et al.* Asymmetric total synthesis of erythromycin. 3. Total synthesis of erythromycin. *J. Am. Chem. Soc.* **103,** 3215–3217 (1981).
6. Martin, S. F., Hida, T., Kym, P. R., Loft, M. & Hodgson, A. The asymmetric synthesis of erythromycin B. *J. Am. Chem. Soc.* **119,** 3193–3194 (1997).
7. Kim, H. C. & Kang, S. H. Total synthesis of azithromycin. *Angew. Chem. Int. Ed.* **48,** 1827–1829 (2009).
8. Andrade, R. B. Total synthesis of desmethyl macrolide antibiotics. *Synlett* **26,** 2199–2215 (2015).
9. Khosla, C. Harnessing the biosynthetic potential of modular polyketide synthases. *Chem. Rev.* **97,** 2577–2590 (1997).
10. Cane, D. E., Walsh, C. T. & Khosla, C. Harnessing the biosynthetic code: combinations, permutations, and mutations. *Science* **282,** 63–68 (1998).
11. Marsden, A. F. A. *et al.* Engineering broader specificity into an antibiotic-producing polyketide synthase. *Science* **279,** 199–202 (1998).
12. Park, S. R. *et al.* Genetic engineering of macrolide biosynthesis: past advances, current state, and future prospects. *Appl. Microbiol. Biotechnol.* **85,** 1227–1239 (2010).
13. Kurath, P., Jones, P. H., Egan, R. S. & Perun, T. J. Acid degradation of erythromycin A and erythromycin B. *Experientia* **27,** 362 (1971).
14. Bright, G. M. *et al.* Synthesis, *in vitro* and *in vivo* activity of novel 9-deoxo-9a-aza-9a-homoerythromycin A derivatives – a new class of macrolide antibiotics, the azalides. *J. Antibiot. (Tokyo)* **41,** 1029–1047 (1988).
15. Morimoto, S., Takahashi, Y., Watanabe, Y. & Omura, S. Chemical modification of erythromycins. 1. Synthesis and antibacterial activity of 6-*O*-methylerythromycins-A. *J. Antibiot. (Tokyo)* **37,** 187–189 (1984).
16. Fernandes, P. B. Methods for treating gastrointestinal diseases. WO patent 2010/048599 (2010).
17. Putnam, S. D., Castanheira, M., Moet, G. J., Farrell, D. J. & Jones, R. N. CEM-101, a novel fluoroketolide: antimicrobial activity against a diverse collection of Gram-positive and Gram-negative bacteria. *Diagn. Microbiol. Infect. Dis.* **66,** 393–401 (2010).
18. Leclercq, R. Mechanisms of resistance to macrolides and lincosamides: nature of the resistance elements and their clinical implications. *Clin. Infect. Dis.* **34,** 482–492 (2002).
19. Gaynor, M. & Mankin, A. S. Macrolide antibiotics: binding site, mechanism of action, resistance. *Curr. Top. Med. Chem.* **3,** 949–961 (2003).
20. Subramanian, S. L., Ramu, H. & Mankin, A. S. in *Antibiotic Discovery and Development* (eds Dougherty, T. J. & Pucci, M. J.) 455–484 (Springer, 2012).
21. Bryskier, A. Ketolides-telithromycin, an example of a new class of antibacterial agents. *Clin. Microbiol. Infect.* **6,** 661–669 (2000).
22. Zhanel, G. G. *et al.* The ketolides: a critical review. *Drugs* **62,** 1771–1804 (2002).
23. Retsema, J. *et al.* Spectrum and mode of action of azithromycin (CP-62,993), a new 15-membered-ring macrolide with improved potency against Gram-negative organisms. *Antimicrob. Agents Chemother.* **31,** 1939–1947 (1987).
24. Girard, A. E. *et al.* Pharmacokinetic and *in vivo* studies with azithromycin (CP-62,993), a new macrolide with an extended half-life and excellent tissue distribution. *Antimicrob. Agents Chemother.* **31,** 1948–1954 (1987).
25. Denis, A. & Agouridas, C. Synthesis of 6-*O*-methyl-azithromycin and its ketolide analog via Beckmann rearrangement of 9(*E*)-6-*O*-methyl-erythromycin oxime. *Bioorg. Med. Chem. Lett.* **8,** 2427–2432 (1998).
26. Boeckman, R. K., Jr & Pruitt, J. R. A new, highly efficient, selective methodology for formation of medium-ring and macrocyclic lactones via intramolecular ketene trapping: an application to a convergent synthesis of (–)-kromycin. *J. Am. Chem. Soc.* **111,** 8286–8288 (1989).
27. Boeckman, R. K. Jr & Perni, R. B. Studies directed toward the synthesis of naturally occurring acyltetramic acids. 2. Preparation of the macrocyclic subunit of ikarugamycin. *J. Org. Chem.* **51,** 5486–5489 (1986).
28. Woodward, R. B. *et al.* Asymmetric total synthesis of erythromycin. 2. Synthesis of an erythronolide A lactone system. *J. Am. Chem. Soc.* **103,** 3213–3215 (1981).
29. Rostovtsev, V. V., Green, L. G., Fokin, V. V. & Sharpless, K. B. A stepwise Huisgen cycloaddition process: copper(I)-catalyzed regioselective "ligation" of azides and terminal alkynes. *Angew. Chem. Int. Ed.* **41,** 2596–2599 (2002).
30. Berisio, R. *et al.* Structural insight into the antibiotic action of telithromycin against resistant mutants. *J. Bacteriol.* **185,** 4276–4279 (2003).
31. Tu, D., Blaha, G., Moore, P. B. & Steitz, T. A. Structures of MLSBK antibiotics bound to mutated large ribosomal subunits provide a structural explanation for resistance. *Cell* **121,** 257–270 (2005).
32. Llano-Sotelo, B. *et al.* Binding and action of CEM-101, a new fluoroketolide antibiotic that inhibits protein synthesis. *Antimicrob. Agents Chemother.* **54,** 4961–4970 (2010).
33. Bulkley, D., Innis, C. A., Blaha, G. & Steitz, T. A. Revisiting the structures of several antibiotics bound to the bacterial ribosome. *Proc. Natl Acad. Sci. USA* **107,** 17158–17163 (2010).
34. Eyal, Z. *et al.* Structural insights into species-specific features of the ribosome from the pathogen *Staphylococcus aureus*. *Proc. Natl Acad. Sci. USA* **112,** E5805–E5814 (2015).
35. Seiple, I. B., Hog, D. T. & Myers, A. G. Practical protocols for the preparation of highly enantioenriched silyl ethers of (*R*)-3-hydroxypentan-2-one, building blocks for the synthesis of macrolide antibiotics. *Synlett* **27,** 57–60 (2016).
36. Seiple, I. B., Mercer, J. A. M., Sussman, R. J., Zhang, Z. & Myers, A. G. Stereocontrolled synthesis of syn-β-hydroxy-α-amino acids by direct aldolization of pseudoephenamine glycinamide. *Angew. Chem. Int. Ed.* **53,** 4642–4647 (2014).
37. Mikami, K., Kawakami, Y., Akiyama, K. & Aikawa, K. Enantioselective catalysis of ketoester-ene reaction of silyl enol ether to construct quaternary carbons by chiral dicationic palladium(II) complexes. *J. Am. Chem. Soc.* **129,** 12950–12951 (2007).
38. Mukaiyama, T., Narasaka, K. & Banno, K. New aldol type reaction. *Chem. Lett.* **2,** 1011–1014 (1973).

© 2016 Macmillan Publishers Limited. All rights reserved

39. Zhang, Z., Kitamura, Y. & Myers, A. G. An efficient directed Claisen reaction allows for rapid construction of 5,6-disubstituted 1,3-dioxin-4-ones. *Synthesis* **47,** 2709–2712 (2015).
40. Hauske, J. R. & Schulte, G. R. Desosamino derivatives of macrolides as immunosuppressants and antifungal agents. WO patent 1993018042 (1993).
41. Zhang, Z., Fukuzaki, T. & Myers, A. G. Synthesis of D-desosamine and analogs by rapid assembly of 3-amino sugars. *Angew. Chem. Int. Ed.* **55,** 523–527 (2016).
42. Leyes, A. E. & Poulter, C. D. Synthesis of (*R*)-[2-^{2}H]isopentenyl diphosphate and determination of its enantiopurity by ^{2}H NMR spectroscopy in a lyotropic medium. *Org. Lett.* **1,** 1067–1070 (1999).
43. Tan, Z. & Negishi, E.-i. An efficient and general method for the synthesis of α,ω-difunctional reduced polypropionates by Zr-catalyzed asymmetric carboalumination: synthesis of the scyphostatin side chain. *Angew. Chem. Int. Ed.* **43,** 2911–2914 (2004).
44. Santaniello, E., Ferraboschi, P. & Grisenti, P. An efficient chemo-enzymatic approach to the enantioselective synthesis of 2-methyl-1,3-propanediol derivatives. *Tetrahedr. Lett.* **31,** 5657–5660 (1990).
45. Thompson, A. S., Corley, E. G., Huntington, M. F. & Grabowski, E. J. J. Use of an ephedrine alkoxide to mediate enantioselective addition of an acetylide to a prochiral ketone: asymmetric synthesis of the reverse transcriptase inhibitor L-743,726. *Tetrahedr. Lett.* **36,** 8937–8940 (1995).
46. Sato, F., Ishikawa, H., Watanabe, H., Miyake, T. & Sato, M. Specific hydromagnesiation of prop-2-ynylic alcohols. A simple and specific route to terpenoids. *J. Chem. Soc. Chem. Commun.* 718–720 (1981).
47. Ashley, G. *et al.* Ketolide antibacterials. US patent 6,590,083 (2003).
48. Antibiotic Resistance Threats in the United States, 2013. http://www.cdc.gov/drugresistance/threat-report-2013/pdf/ar-threats-2013-508.pdf (CDC, US Dept Health and Human Services, 2013).
49. Charest, M. G., Lerner, C. D., Brubaker, J. D., Siegel, D. R. & Myers, A. G. A convergent enantioselective route to structurally diverse 6-deoxytetracycline antibiotics. *Science* **308,** 395–398 (2005).
50. Liu, F. & Myers, A. G. Development of a platform for the discovery and practical synthesis of new tetracycline antibiotics. *Curr. Opin. Chem. Biol.* **32,** 48–57 (2016).

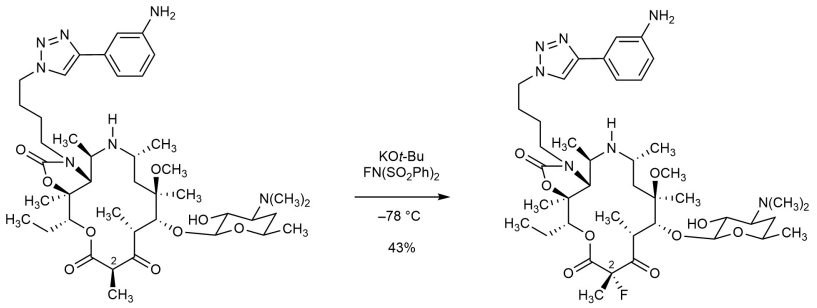
Supplementary Information is available in the online version of the paper.

Acknowledgements We acknowledge funding from Alistair and Celine Mactaggart, from the Gustavus and Louise Pfeiffer Research Foundation, and from the Blavatnik Biomedical Accelerator Program at Harvard University. We thank NERCE (NIH project number U54 AI057159), W. Weiss (Univ. North Texas), and R. Alm and S. Lahiri (Macrolide Pharmaceuticals) for measuring MIC values, R. Alm and T. Dougherty (Harvard Medical School) for genetic characterization of a microbial resistance gene, and S.-L. Zheng (Harvard University) for conducting X-ray crystallographic analyses. I.B.S. acknowledges postdoctoral fellowship support from the National Institutes of Health (F32GM099233); Z.Z. is a Howard Hughes Medical Institute International Student Research fellow; A.L.-M. acknowledges postdoctoral fellowship support from the Swiss National Science Foundation (PBGEPE2-139864) and the Novartis Foundation; D.T.H. is indebted to the Deutsche Forschungsgemeinschaft (DFG) for a postdoctoral fellowship (HO 5326/1-1); T.F. acknowledges Daiichi-Sankyo Co., Ltd, for financial support; and Y.K. acknowledges support from the Engineering Promotion Fund of Gifu University.

Author Contributions I.B.S., Z.Z. and A.G.M. identified the targets and designed the synthetic routes; I.B.S. and Z.Z. executed and optimized the synthetic routes described in the main text; I.B.S., Z.Z., P.J., P.M.W., A.L.-M. and D.T.H. executed the synthetic routes shown in the Extended Data; I.B.S., Z.Z., P.J., P.M.W., A.L.-M., D.T.H., K.Y., T.F., P.N.C., X.Z., M.L.C., F.T.S. and W.D.G. synthesized individual macrolide analogues; I.B.S., Z.Z., Y.K. and S.R.A. synthesized and scaled the building blocks. I.B.S., Z.Z. and A.G.M. wrote the paper. All authors discussed the results and commented on the manuscript.

Author Information Atomic coordinates and structure factors for the crystal structure reported have been deposited with the Cambridge Crystallographic Database under accession number CCDC 1440650. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.G.M. (myers@chemistry.harvard.edu).

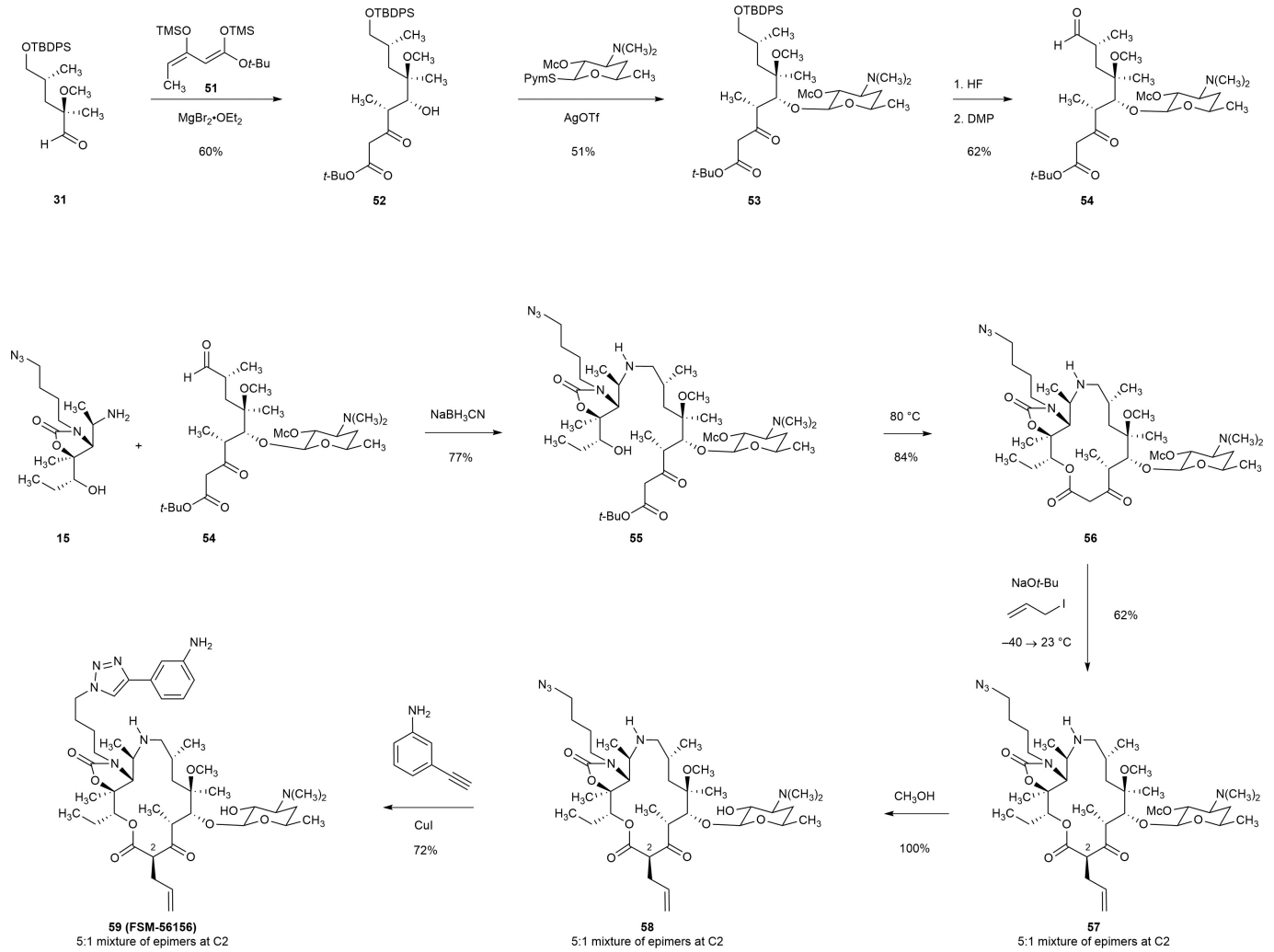
© 2016 Macmillan Publishers Limited. All rights reserved



25 (FSM-22367) 50 (FSM-22391)

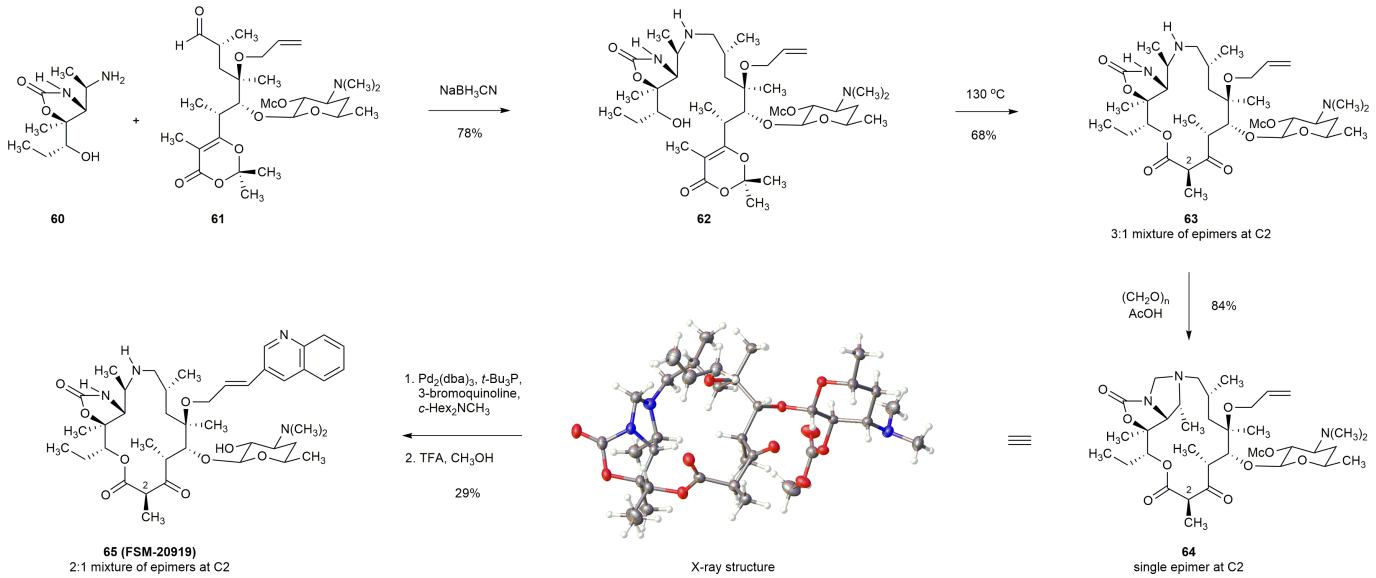
Extended Data Figure 1 | Synthesis of a C2-fluoro 14-membered azaketolide by a late-stage fluorination reaction. Subjection of β-keto lactone **25** (FSM-22367) to potassium *tert*-butoxide (1.0 equiv.) at −78 °C followed by *N*-fluorobenzenesulfonimide (1.0 equiv.) afforded **50** (FSM-22391) in 43% yield.

© 2016 Macmillan Publishers Limited. All rights reserved



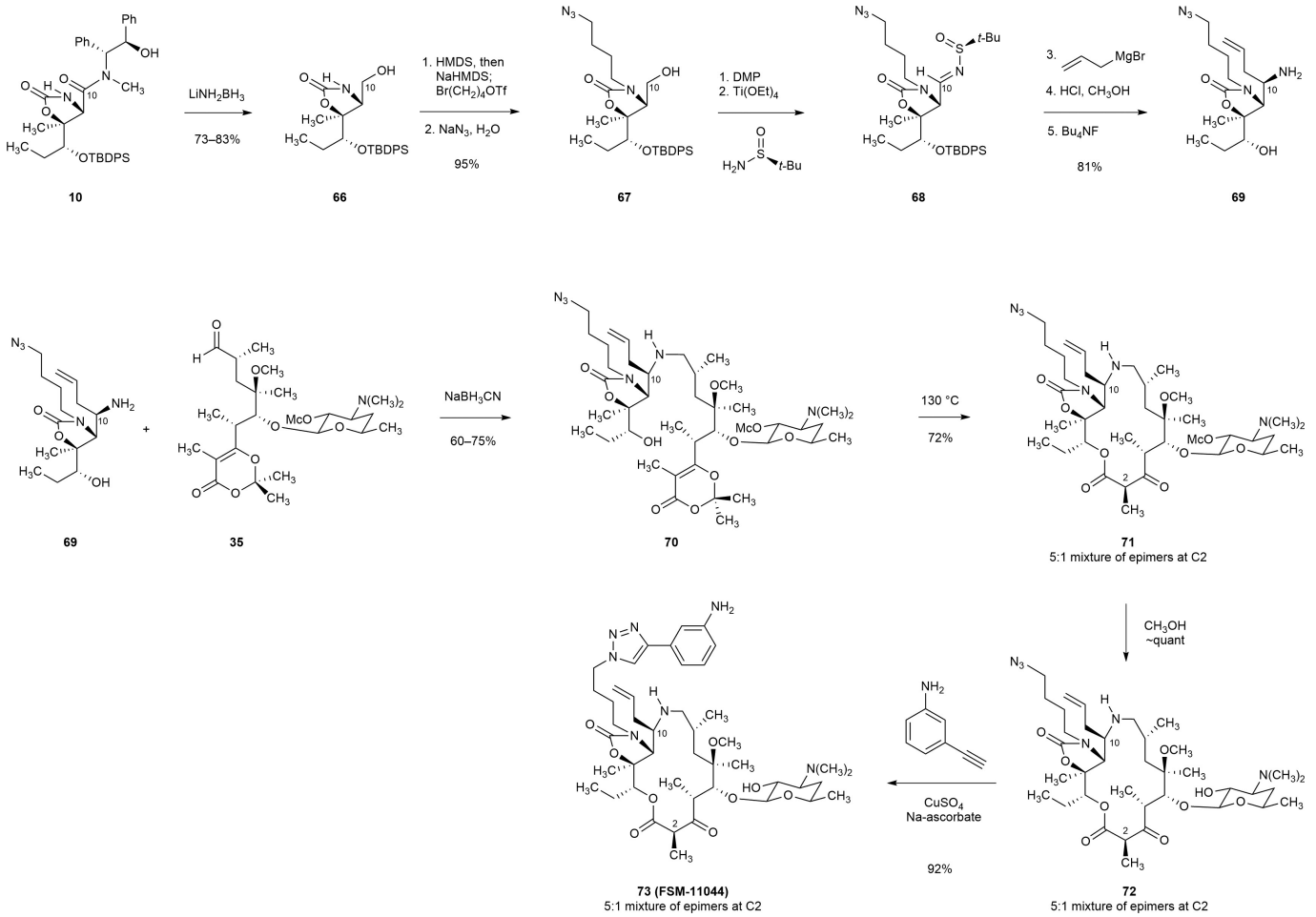
Extended Data Figure 2 | Synthesis of a 15-membered azaketolide with a modified C2-substituent. Thermolysis of a β-keto *tert*-butyl ester substrate (**55**) proceeded at a lower temperature (80 °C) than that for dioxolenone substrates (132 °C) and afforded a 15-membered macrocycle without substitution at C2 (**56**). (Here and elsewhere we omit mention of the intermediate compounds for clarity.) This macrocycle served as a nearly ideal intermediate for preparation of macrolides with diverse C2-substitutions. For example, an allyl group was introduced at C2 by treatment of **56** with sodium *tert*-butoxide (1.1 equiv.) and allyl iodide (1.1 equiv.) at −40 °C followed by warming the reaction solution to 23 °C. The product **57** (obtained in 62% yield) was then transformed to **59** (FSM-56156) in two steps (via **58**; 72% yield).

© 2016 Macmillan Publishers Limited. All rights reserved



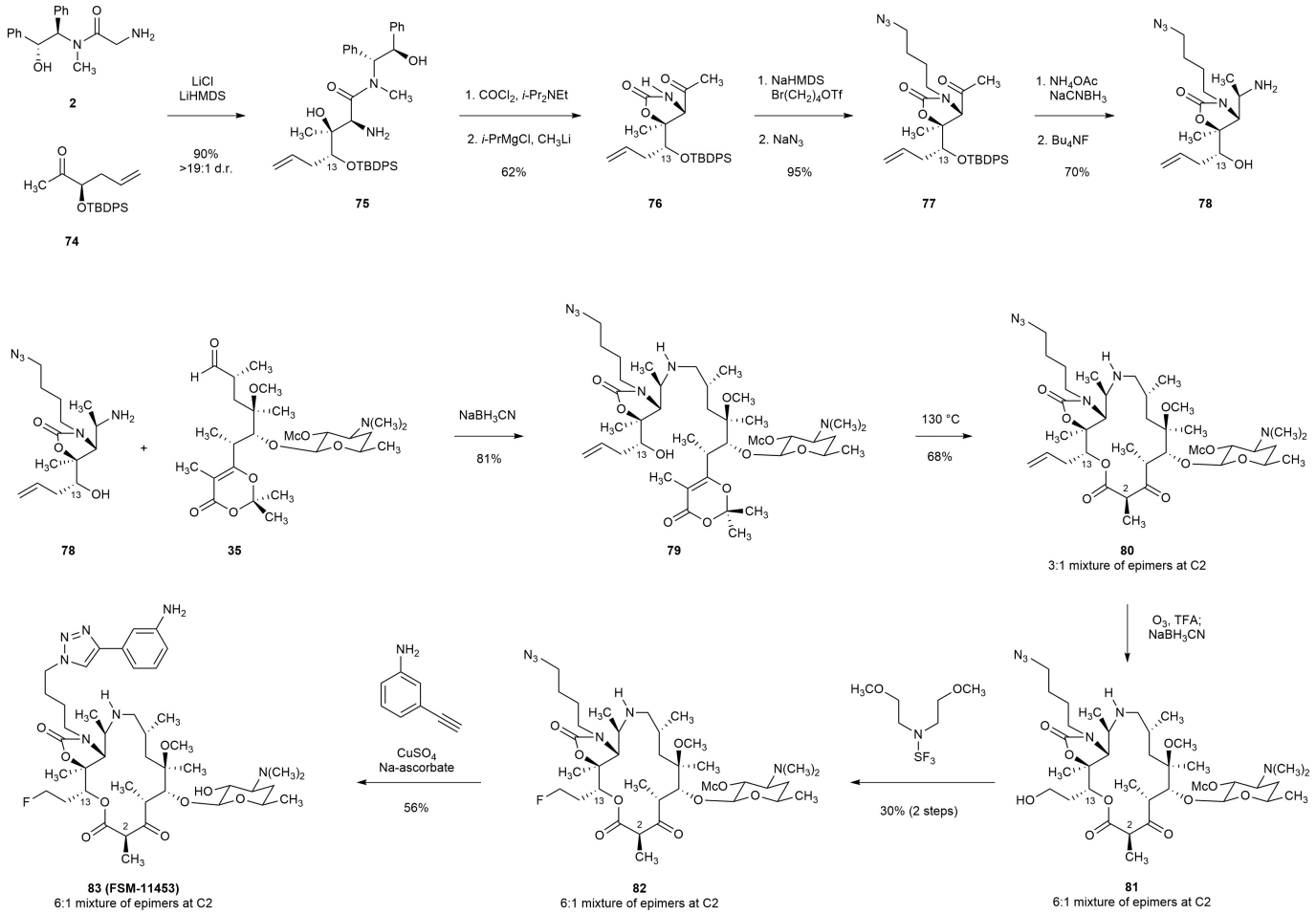
Extended Data Figure 3 | Synthesis of a 15-membered azacethromycin hybrid. Macrolactone **63** was prepared from amine **60** and aldehyde **61** in two steps (via **62**) by a reductive amination–macrocyclization sequence. Treatment of **63** with paraformaldehyde (6.0 equiv.) and acetic acid (10.0 equiv.) furnished adduct **64** as a crystalline solid (84% yield; X-ray structure shown). The imidazolidine group within **64** served to protect both the secondary amine and the cyclic carbamate functions, and permitted the introduction of a quinoline heterocycle via a Heck reaction. Methanolysis (TFA, CH_3OH) cleaved the imidazolidine group, affording **65** (FSM-20919; 29%, two-step yield).

© 2016 Macmillan Publishers Limited. All rights reserved



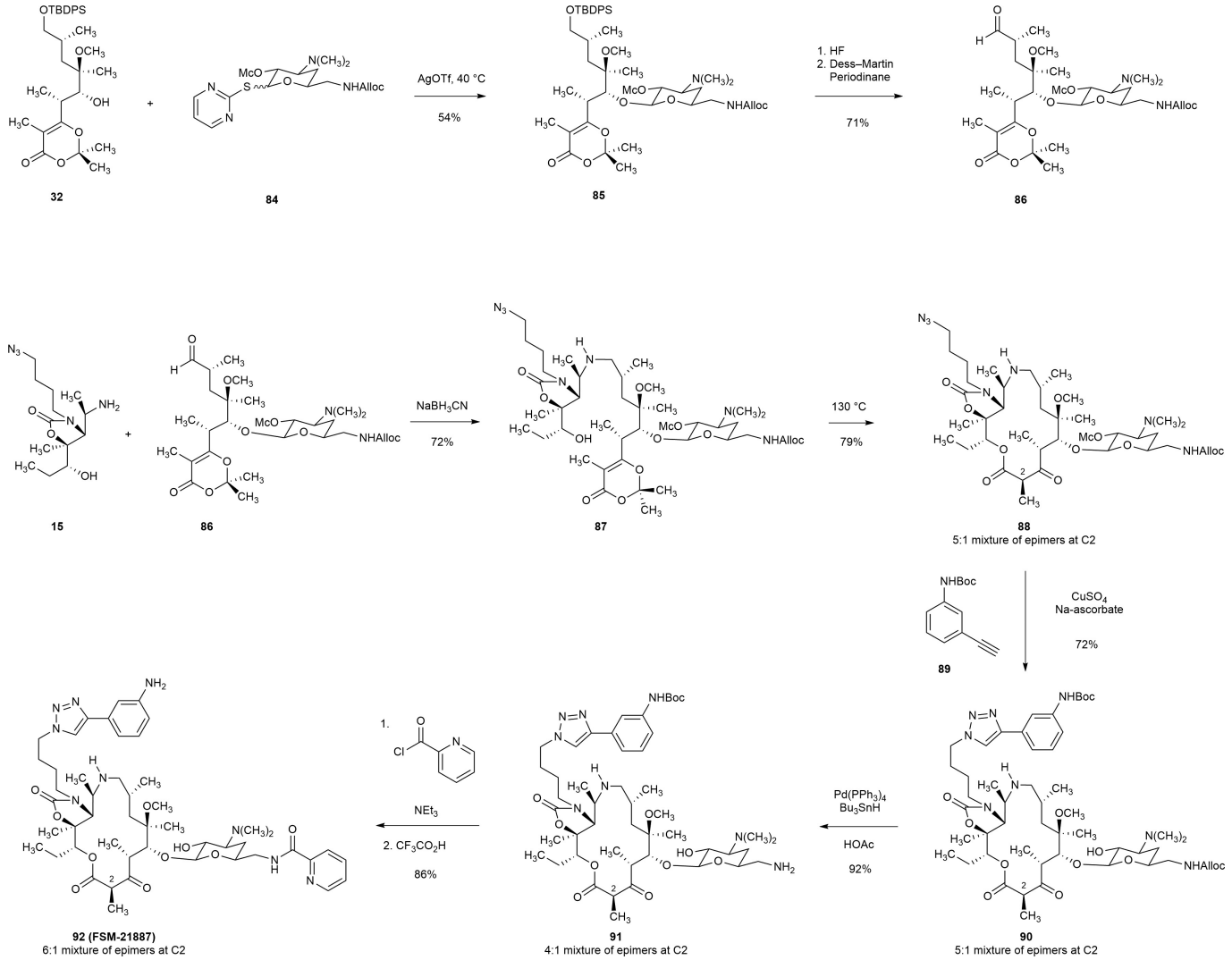
Extended Data Figure 4 | Synthesis of a 15-membered azaketolide with a modified C10-substituent. *N-tert*-butylsulfinyl imine **68** (prepared in five steps via **66** and **67** from amide **10**) allowed for the stereocontrolled introduction of various C10-substituents. For example, addition of allylmagnesium bromide proceeded with >20:1 stereoselectively to establish the stereocentre at C10; subsequent cleavage of the sulfinyl (HCl, CH_3OH) and *tert*-butyldiphenylsilyl (Bu_4NF) groups within the adduct then furnished left-hand intermediate **69** (81% yield). Amine **69** and aldehyde **35** were coupled by a reductive amination reaction ($NaBH_3CN$, 60%–75% yield). The product (**70**) was then transformed to **73** (FSM-11044) in a three-step sequence that consisted of a macrocyclization reaction (giving **71**; 72% yield), a methanolysis reaction (giving **72** in quantitative yield) and lastly a [3 + 2] dipolar cycloaddition reaction (giving **73** in 92% yield).

© 2016 Macmillan Publishers Limited. All rights reserved



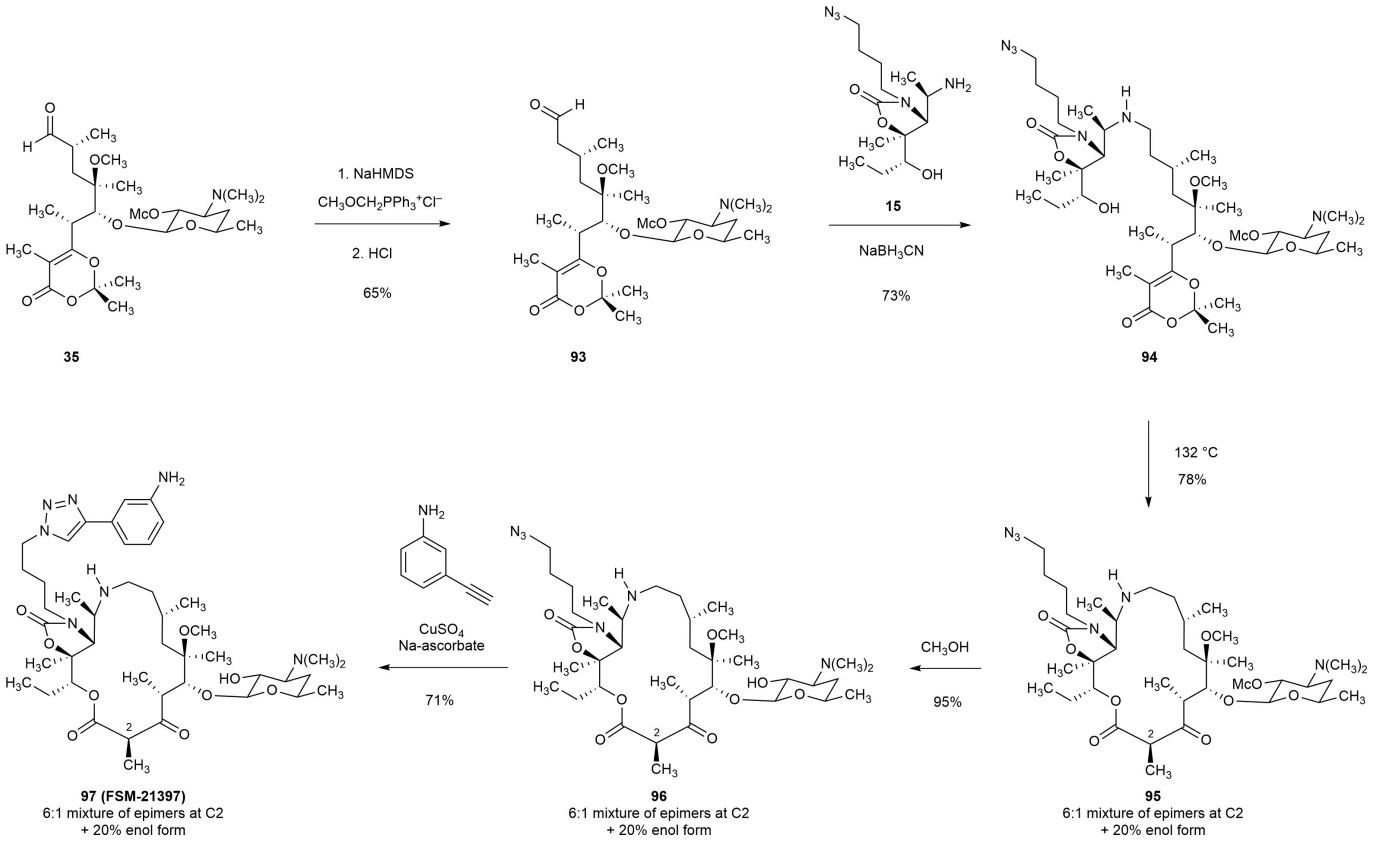
Extended Data Figure 5 | Synthesis of a 15-membered azaketolide with a modified C13-substituent. Modification of position C13 was achieved by modification of a single component, in this case the ketone building block **74** depicted above. Reductive coupling of **78** and **35** united the left- and right-halves to give **79**; subsequent thermal macrocyclization provided macrolactone **80**. The allyl group within intermediate **80** was cleaved upon ozonolysis (O_3, trifluoroacetic acid); reductive workup with sodium cyanoborohydride afforded alcohol **81**. Subjection of **81** to bis(2-methoxyethyl)aminosulfur trifluoride afforded the fluoroethyl-substituted macrocycle **82** (30%, two-step yield), which was transformed to **83** (FSM-11453) by a [3 + 2] dipolar cycloaddition reaction.

© 2016 Macmillan Publishers Limited. All rights reserved



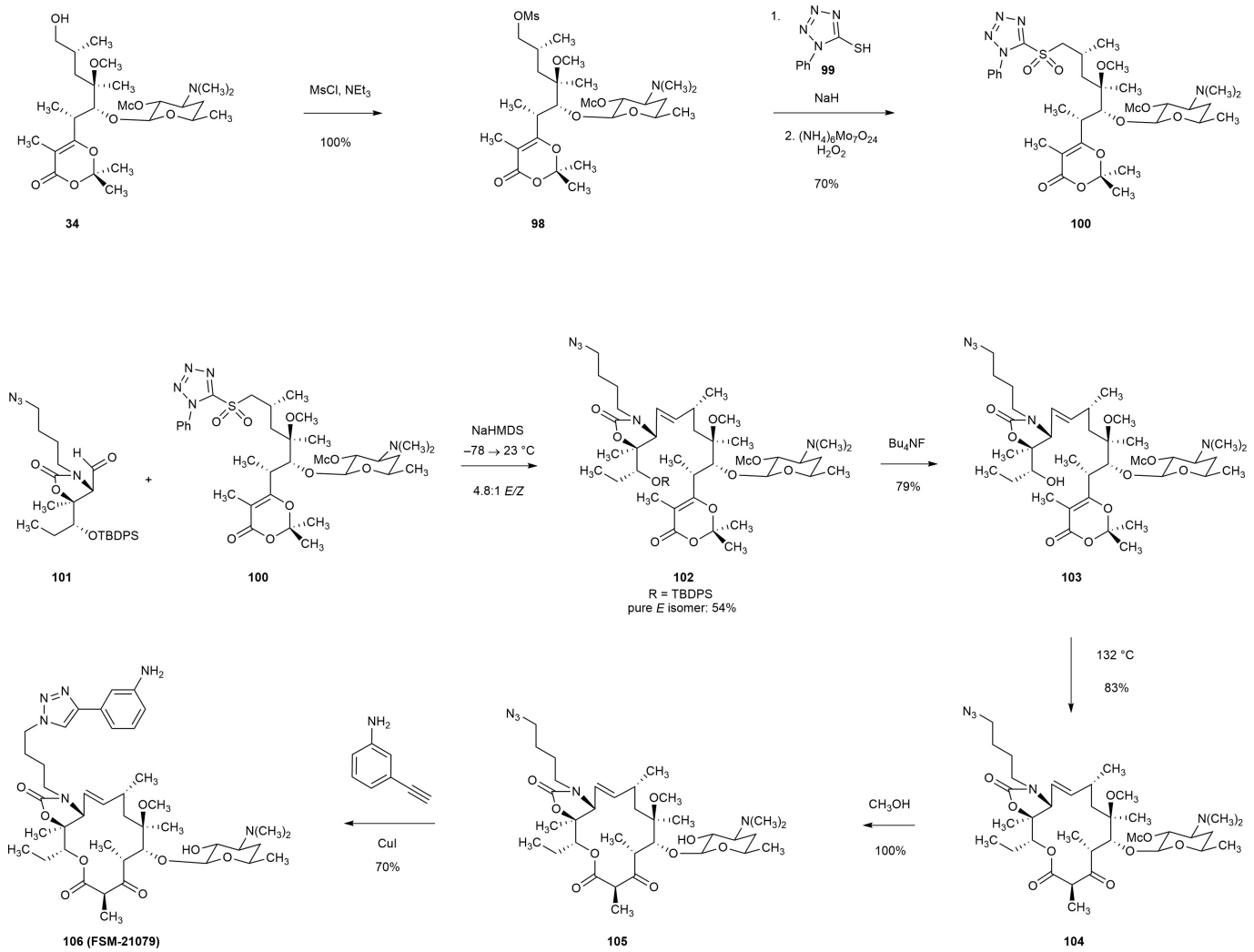
Extended Data Figure 6 | Synthesis of a 15-membered azaketolide with a modified desosamine sugar residue. The 15-membered macrolactone **90** was synthesized using thioglycoside **84** and alkyne **89** as building blocks (in lieu of building blocks **28** and **8** used for the synthesis of 15-membered azaketolide FSM-20707). Treatment of **90** with tributyltin hydride (2.0 equiv.), acetic acid (5.0 equiv.), and tetrakis(triphenylphosphine)palladium (2 mol%) led to cleavage of the allyloxycarbonyl protective group, giving rise to amine **91** (92% yield). The latter intermediate has been transformed into a number of fully synthetic macrolides with modified desosamine sugar residues. For example, acylation of the primary amino group of intermediate **91** with pyridine-2-carbonyl chloride (2.0 equiv.) in the presence of trimethylamine (3.0 equiv.) followed by removal of the *tert*-butoxycarbonyl group afforded **92** (FSM-21887, 86%, two-step yield).

© 2016 Macmillan Publishers Limited. All rights reserved



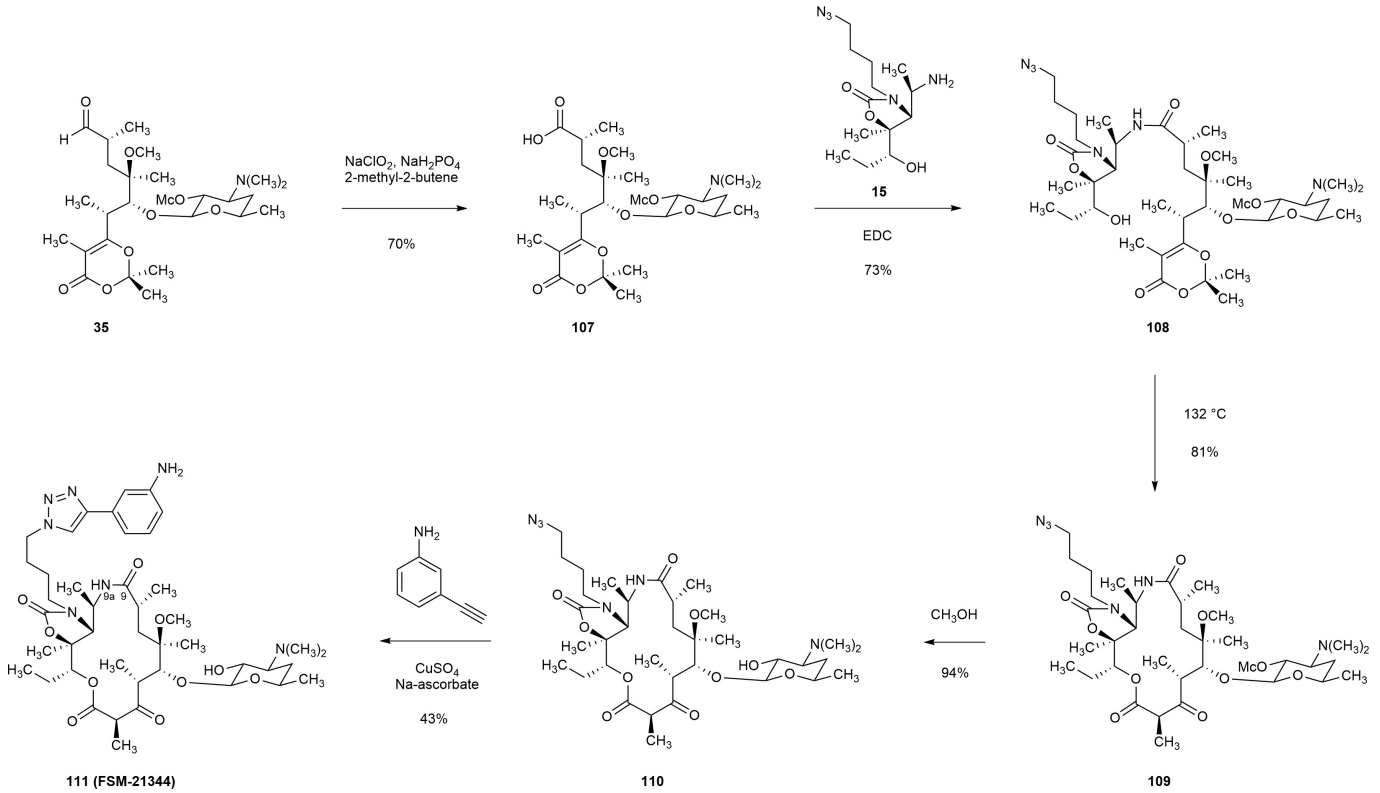
Extended Data Figure 7 | Synthesis of a 16-membered azaketolide. Homologation of aldehyde **35** was achieved by a Wittig olefination reaction ($CH_3OCH_3PPh_3^+$ Cl^-, NaHMDS) followed by hydrolysis of the resulting enol ether to afford aldehyde **93** in 65% yield. Reductive coupling of amine **15** and aldehyde **93** furnished macrocyclization precursor **94** (73% yield). The 16-membered macrolactone **95** was obtained in 78% yield upon thermolysis of **94** (1 mM, 132 °C). Two additional steps transformed **95** via **96** to the 16-membered azaketolide **97** (FSM-21397).

© 2016 Macmillan Publishers Limited. All rights reserved



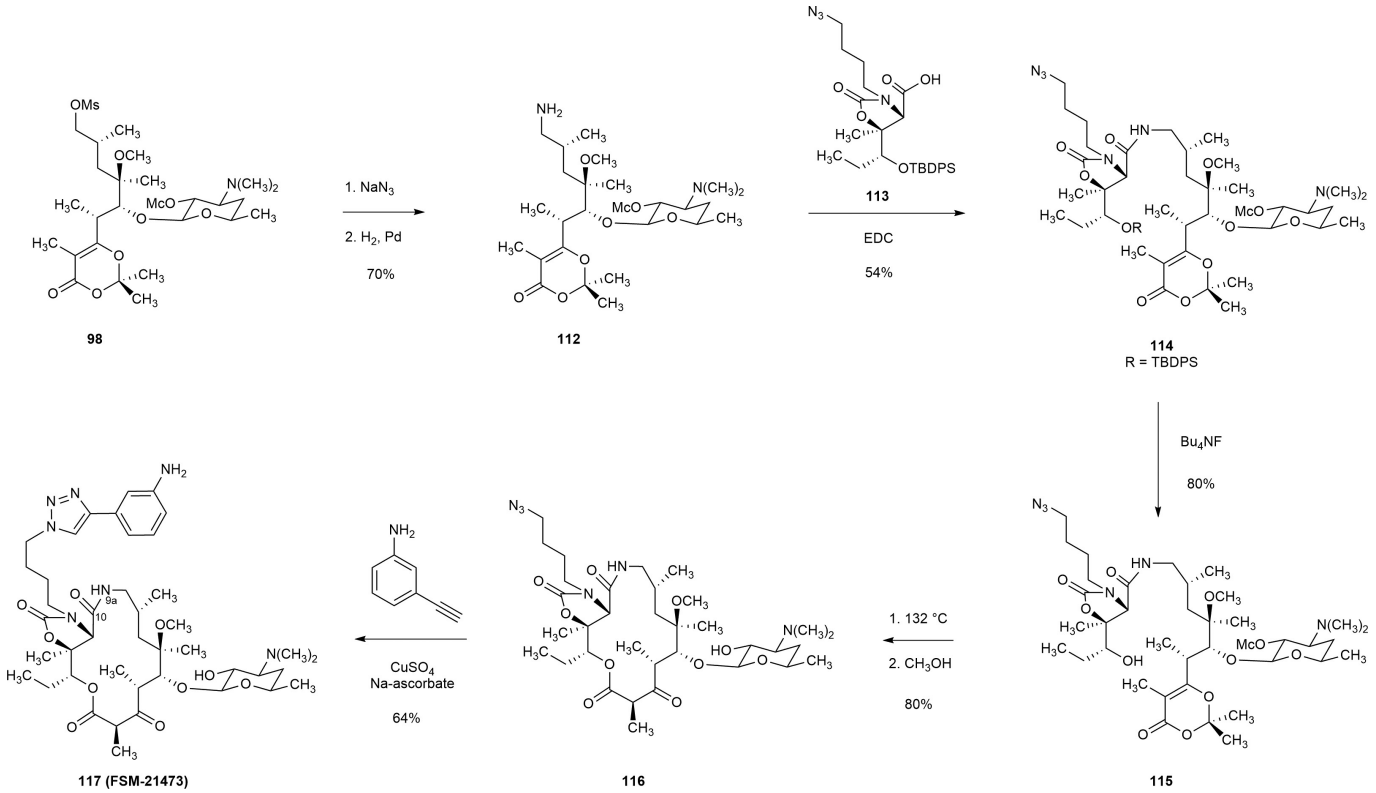
Extended Data Figure 8 | Synthesis of a 14-membered macrolide with a *trans*-olefin linkage. Mesylate **98** was prepared in quantitative yield by treatment of alcohol **34** with methanesulfonyl chloride (1.50 equiv.) and triethylamine (2.0 equiv.). Displacement of the mesylate group in **98** with sodium 1-phenyl-*1H*-tetrazole-5-thiolate (2.0 equiv.) followed by oxidation of the resulting thioether with ammonium molybdate (0.20 equiv.)–hydrogen peroxide (100 equiv.) afforded sulfone **100** in 70% yield. Aldehyde **101** and sulfone **100** were coupled in a Julia–Kocienski olefination reaction (NaHMDS, −78 → 23 °C) to provide a 4.8:1 mixture of *E*- and *Z*-olefin isomers. The *E*-isomer **102** was isolated and desilylated (Bu_4NF, 79%). Thermolysis of the product **103** (1 mM, 132 °C) furnished the 14-membered macrocycle **104** in 83% yield. **104** was then transformed to **106** (FSM-21079) in two additional steps (methanolysis and [3 + 2] dipolar cycloaddition) via **105**.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 9 | Synthesis of a 15-membered macrolide with an amide linkage (C9-N9a). Oxidation of aldehyde **35** with sodium chlorite (10.0 equiv.) in the presence of sodium dihydrogen phosphate (10.0 equiv.) and 2-methyl-2-butene (100 equiv.) afforded carboxylic acid **107** in 70% yield. Acid **107** and amine **15** were coupled in the presence of 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide (2.00 equiv.) to provide amide **108**. Macrocyclization of **108** (1 mM, 132 °C) proceeded in 81% yield to afford macrolactam **109**. Methanolysis and [3 + 2] dipolar cycloaddition then transformed **109** to **111** (FSM-21344) in two steps via **110**.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 10 | Synthesis of a 15-membered macrolide with an amide linkage (C10-N9a). Amine **112** was prepared in 70% yield by displacement of the mesylate group in **98** with sodium azide followed by reduction of the resulting alkyl azide (H_2, Pd). The coupling of amine **112** and acid **113** proceeded in 54% yield. The product, amide **114**, was desilylated (Bu_4NF, 80%) to afford the macrocyclization precursor **115**. Thermal macrocyclization of **115** followed by cleavage of the methoxycarbonyl protective group afforded lactam **116** in 80% yield. Copper-catalysed [3 + 2] dipolar cycloaddition then provided **117** (FSM-21473) in 64% yield.

© 2016 Macmillan Publishers Limited. All rights reserved

ARTICLE

doi:10.1038/nature17971

Structure of the T4 baseplate and its function in triggering sheath contraction

Nicholas M. I. Taylor[1], Nikolai S. Prokhorov[1,2], Ricardo C. Guerrero-Ferreira[1], Mikhail M. Shneider[1,3], Christopher Browning[1]†, Kenneth N. Goldie[4], Henning Stahlberg[4] & Petr G. Leiman[1]

Several systems, including contractile tail bacteriophages, the type VI secretion system and R-type pyocins, use a multiprotein tubular apparatus to attach to and penetrate host cell membranes. This macromolecular machine resembles a stretched, coiled spring (or sheath) wound around a rigid tube with a spike-shaped protein at its tip. A baseplate structure, which is arguably the most complex part of this assembly, relays the contraction signal to the sheath. Here we present the atomic structure of the approximately 6-megadalton bacteriophage T4 baseplate in its pre- and post-host attachment states and explain the events that lead to sheath contraction in atomic detail. We establish the identity and function of a minimal set of components that is conserved in all contractile injection systems and show that the triggering mechanism is universally conserved.

Bacteriophages (viruses of bacteria) use a specialized organelle called a tail to deliver their genetic material and proteins across the cell envelope during infection[1–4]. In phages with the most complex contractile tails, attachment to the host cell is accompanied by a substantial transformation of the tail structure: the external tail sheath contracts and drives a spike-tipped[5,6], rigid tube through the host cell membrane. Other macromolecular complexes, such as the type VI secretion system (T6SS)[7–11], metamorphosis-associated contractile (MAC) arrays[12], R-type pyocins[13,14], *Serratia* antifeeding prophage[15], *Photorhabdus* virulence cassette[16] and rhapidosomes[1,17], use a similar contractile sheath–rigid tube mechanism to breach the bacterial or eukaryotic cell envelope. However, the most complex part of these 'contractile injection systems'[18]—the baseplate, which is responsible for coordinating host recognition or other environmental signals with sheath contraction—remains poorly characterized, and even its exact protein composition remains unknown for otherwise well studied systems such as bacteriophage T4 (refs 1, 8).

The T4 baseplate is currently thought to contain at least 15 different proteins with copy numbers ranging from 1 to 18 (ref. 19). Assembly of the T4 baseplate involves two large independent intermediates: a hub and a wedge[20–22]. Several phage and possibly host cell chaperones mediate the joining of six wedges to the hub, which is then followed by the attachment of receptor-binding fibres to this structure. The nascent baseplate initiates tube assembly and subsequent polymerization of the sheath in the extended high-energy state[23,24].

The remarkable structure and transformation of the T4 tail and other contractile injection systems have received considerable attention[2–5,10,12,15,25–32]. However, many questions remain unanswered, such as how the baseplates are organized at the molecular level, how they circularize, how the tail fibre network communicates with the central part of the baseplate, and how the signal of host cell surface binding leads to the conformational change that releases the tube and induces sheath contraction. To answer these questions, we used cryo-electron microscopy (cryo-EM) to determine the structure of the T4 baseplate–tail tube complex at a resolution of 3.8 Å for the universally conserved central part and tail tube and 4.1 Å overall, and the structure of the post-attachment baseplate at a resolution of 6.8 Å. We have built and refined the atomic model of the baseplate in its pre- and post-host attachment conformations, allowing us to visualize the transformation of this 440,000-atom (not counting hydrogens) structure. We have analysed the signal transduction network of the baseplate by mutagenesis and examined the baseplate assembly pathway in the T6SS.

Overall structure of the T4 baseplate

We produced the pre-host attachment (also known as 'hexagonal'[33]) baseplate–tail tube complex *in vivo* by stopping phage assembly at the stage of sheath polymerization[26] (Extended Data Fig. 1). The complex switched to a post-attachment (also known as 'star-shaped'[33]) conformation upon interaction with the cell membrane and sometimes spontaneously; at this point it lost the tube and central spike complex (Extended Data Fig. 1).

The large size of the T4 baseplate (the diameter is 490 Å, about twice that of the ribosome) made particle alignment very accurate but contributed to a substantial difference in resolution at the centre and edges of the reconstructed volume (Extended Data Fig. 2). We used local mask-based particle alignment[34,35] to improve the electron density around the periphery of the baseplate (Extended Data Fig. 2). Furthermore, because this region is made of mostly trimeric proteins, we were able to improve the maps by additional local three-fold averaging.

The final atomic model is around 96% complete and contains 56,082 amino acid residues (9,886 unique) (Supplementary Tables 1 and 2). These amino acids belong to 145 polypeptide chains of 15 different proteins (gene products, gps) that comprise the baseplate (gp5, gp5.4, gp6, gp7, gp8, gp9, gp10, gp11, gp12, gp25, gp27 and gp53) and the proximal region of the tail tube (gp19, gp48 and gp54, although gp48 and gp54 could also be considered to be baseplate components[19]; Fig. 1). The tape measure protein gp29, which is located inside the tube and is thought to span its entire length[19], is not resolved in our cryo-EM map. Gp5, gp5.4 and gp27 form the central spike complex that creates an extension of the tube. The rest of the baseplate can be divided into inner (gp6, gp25, gp53 and part of gp7), intermediate (gp8 and the

[1]École Polytechnique Fédérale de Lausanne (EPFL), BSP-415, 1015 Lausanne, Switzerland. [2]Winogradsky Institute of Microbiology, Research Center of Biotechnology of the Russian Academy of Sciences, pr. 60-letiya Oktyabrya, 7 build. 2, 117312, Moscow, Russia. [3]Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Laboratory of Molecular Bioengineering, 16/10 Miklukho-Maklaya St., 117997 Moscow, Russia. [4]Center for Cellular Imaging and NanoAnalytics (C-CINA), Biozentrum, University of Basel, Mattenstrasse 26, 4058 Basel, Switzerland.
†Present address: Vertex Pharmaceuticals (Europe) Ltd, 86–88 Jubilee Avenue, Milton Park, Abingdon, Oxfordshire OX14 4RW, UK.

© 2016 Macmillan Publishers Limited. All rights reserved

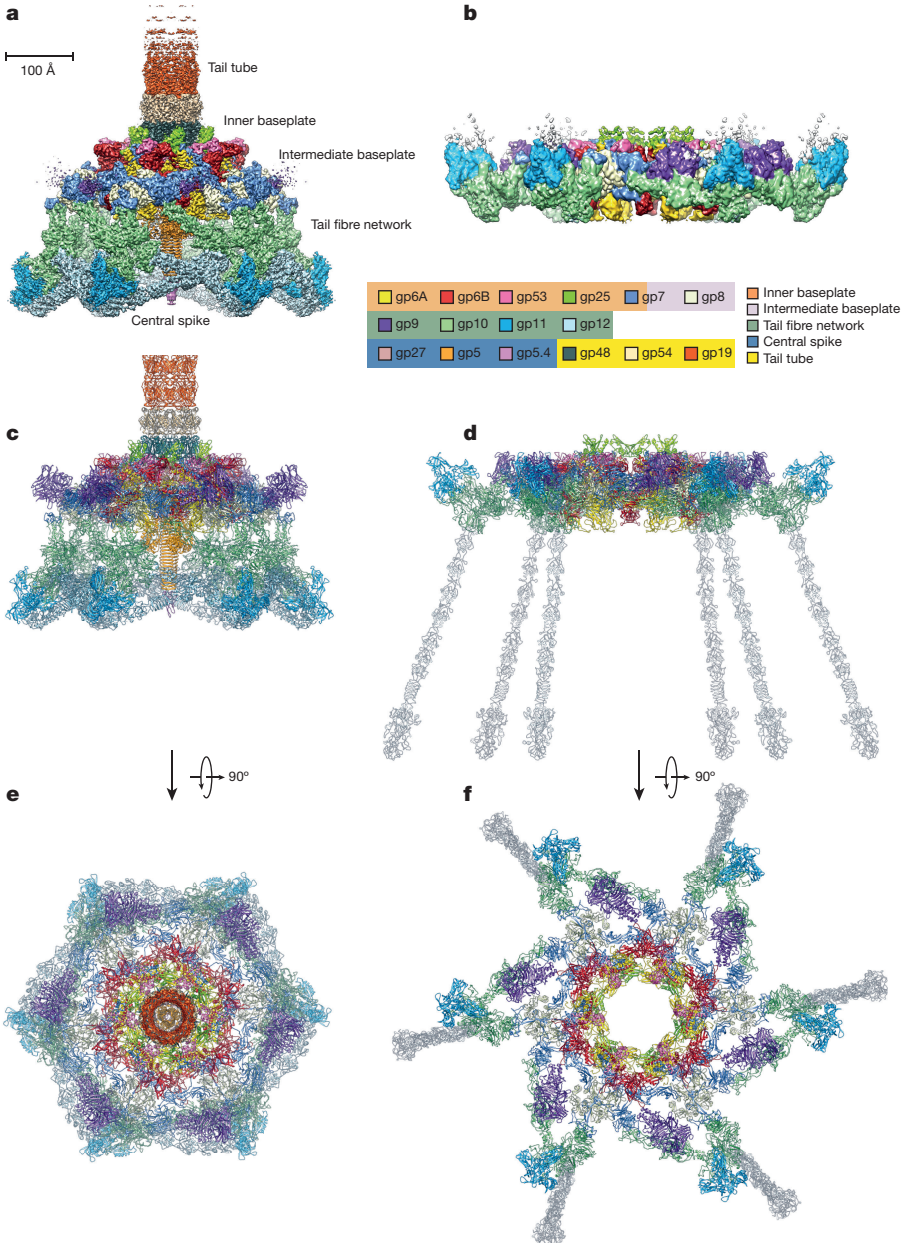


Figure 1 | Maps and atomic models of T4 baseplate in pre- and post-attachment states. a, **b**, Cryo-EM reconstructions of the pre- and post-attachment T4 baseplate, respectively. **c**–**f**, Atomic models of the two states with component proteins shown as ribbon diagrams. In **d** and **f**, the STFs (gp12 trimers) are displayed semi-transparently to indicate that they are not present in the refined model of the post-attachment baseplate.

rest of gp7), and peripheral (gp9, gp10, gp11 and gp12) parts with the peripheral part constituting the tail fibre network (Fig. 1).

The inner baseplate

Twelve copies of gp6 form a continuous ring that makes up most of the inner baseplate (Extended Data Fig. 3). The asymmetric unit of the ring contains two gp6 molecules that are chemically identical but have markedly different conformations. One molecule (gp6A) lies closer to the six-fold axis whereas the other (gp6B) is further away (Extended Data Fig. 3a, b). Each gp6 molecule in the ring interacts with its two neighbours, forming an N-terminal dimer with one and a C-terminal dimer with the other (Extended Data Fig. 3c, d). An earlier study proposed a similar organization of the gp6 ring[36], and the crystal structure reported in that study matches the gp6 C-terminal dimer in the complete baseplate (Extended Data Fig. 3d). However, the interaction of the gp6 N-terminal regions is more complex than presumed earlier.

The N-terminal half of gp6 is part of a $(gp6)_2$–gp7 heterotrimeric module that consists of an α-helical 'core bundle' and a flat 'trifurcation unit' (Fig. 2). Gp25 and gp53 connect the core bundle to the central hub and the tube (Figs 2a, 3a), whereas the trifurcation unit allows the three polypeptide chains comprising it to reach different destinations (Fig. 2d). Two gp6 molecules extend in opposite directions to form C-terminal dimers that connect the trifurcation units into an iris-like structure (Fig. 2d). Gp7 follows a path towards the periphery of the baseplate, where it interacts with the tail fibre network proteins gp9 and gp10. The $(gp6)_2$–gp7 module plays a critical role in the assembly and function of the baseplate because it mediates both the circularization of the baseplate and the communication of the tail fibre network with the baseplate hub.

The $(gp6)_2$–gp7 heterotrimeric module contains large insertions that do not distort the gp6–gp7 iris (Fig. 2d). A wing-shaped domain (residues 101–315) is inserted between the core bundle and the trifurcation unit in gp6 (Fig. 2b). The wing domain of gp6A tightly cradles the hub,

© 2016 Macmillan Publishers Limited. All rights reserved

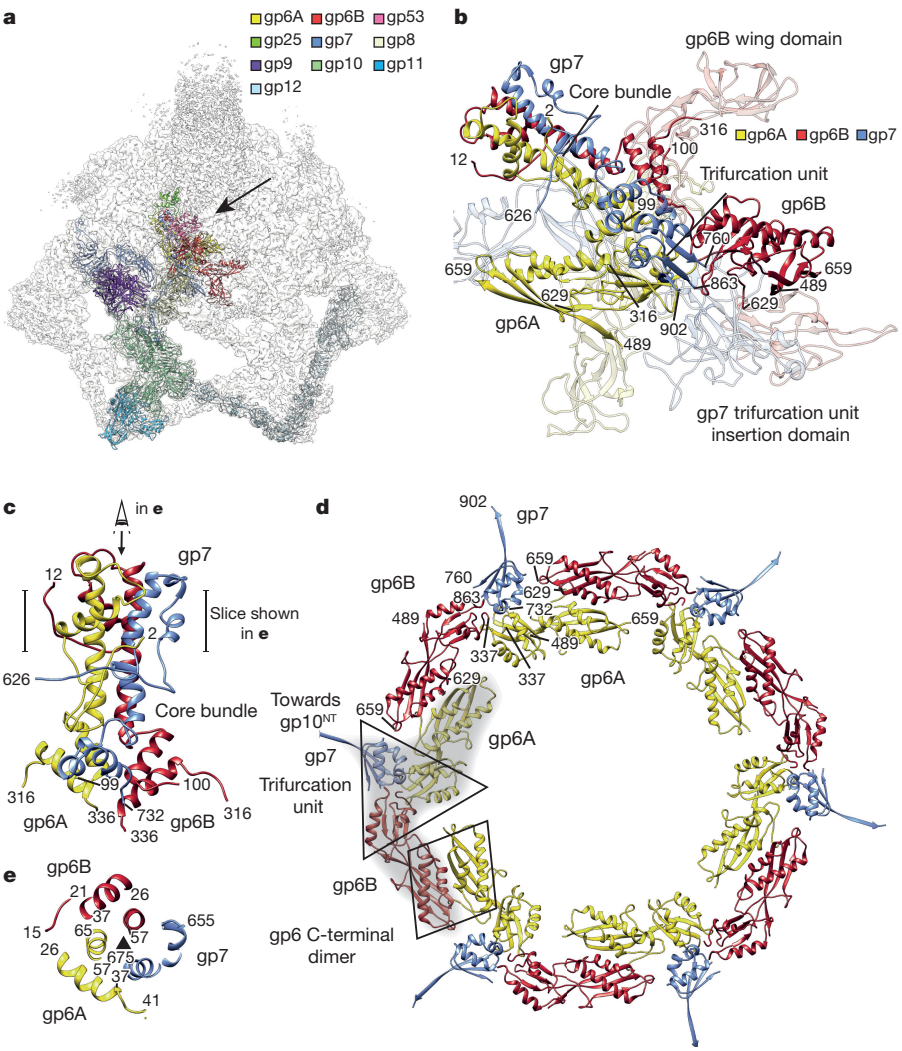


Figure 2 | The $(gp6)_2$–gp7 heterotrimeric unit. a, Composition of the largest assembly intermediate of the baseplate (the wedge) and its location in the cryo-EM map. Gp12 and gp9 are not part of the wedge and are shown as visual clues to help with interpretation. The arrow points to the $(gp6)_2$–gp7 heterotrimeric unit. **b**, The structure of the $(gp6)_2$–gp7 heterotrimer. Regions with lower conservation are shown semi-transparently. **c**, A close-up view of the $(gp6)_2$–gp7 core bundle. **d**, The iris created by the $(gp6)_2$–gp7 trifurcation domains and gp6 C-terminal dimers. **e**, A slice of the structure shown in **c** demonstrates three-fold symmetry.

whereas that of gp6B forms a segment of the platform for assembly of the first layer of the sheath (Extended Data Fig. 3c). In gp7, an insertion domain (residues 761–862) is part of the trifurcation unit. It is positioned close to the three-fold axis of the unit on the side opposite to the core bundle (Fig. 2b).

Gp25 and gp53 form extensive interactions with the core bundle that stabilize the entire assembly. Gp25 is firmly attached to the tip of the bundle (the buried surface area is 1,525 Å^2) with its residues 1–28 capping the bundle (Fig. 3b). Gp53 consists of two domains, one of which (residues 1–103 and 155–193) clamps the core bundle and encircles it for about three-quarters of a turn (the buried surface area is 3,179 Å^2) while the other extends the sheath assembly platform formed by the gp6B wing domain.

Both gp25 and gp53 contain conserved motifs that interact with each other and with other components of the core bundle. The most conserved region of gp53 and its orthologues is a LysM motif[37] (residues 46–82; Fig. 3b). The tip of a structurally conserved loop of gp25 (residues 47–49) mediates the interaction between LysM and the core bundle. Orthologues of gp25 across many contractile ejection systems contain an EPR motif (Glu-Pro-Arg, residues 85–87 of gp25)[38]. This motif interacts with the core bundle (Fig. 3b) and points towards the region of the gp27–gp48 interface, which is not resolved in our cryo-EM map. The remarkable degree of conservation of this motif and its location at the baseplate–tail tube interface suggest that it might perform an additional, possibly enzymatic, function, such as maturational proteolysis, which is widespread in virus assembly[39].

The central spike and the tube

The density of the trimeric part of the $(gp5.4)_1$–$(gp5)_3$–$(gp27)_3$ central spike complex is a superposition of its two non-equivalent positions related by a 60° rotation around the six-fold axis of the baseplate. Nevertheless, the superimposed β-strands of the tandem tube domains of gp27 are perfectly resolved, confirming its three-to-six-fold symmetry adaptor function, as proposed in an earlier study[5]. Thanks to the prominent lysozyme domain of gp5, placement of the trimeric part of the spike in its two possible positions is unambiguous.

The tail tube is a smooth continuation of the spike complex, and gp48 and gp54 form its first two annuli (Fig. 3a). The tube body, as expected[23], is made of stacked gp19 hexamers. All the three proteins have a fold common to tube proteins from other contractile injection systems (for example, T6SS Hcp (ref. 11) or PA0623 of R-type pyocin[14]). The electron density is of sufficient quality to build the atomic model of the first two gp19 rings, allowing us to define the symmetry of the tube. Its helical rise and twist are 40.2 Å and 17.9°, respectively. These parameters are very close to those of the extended sheath determined in previous, lower-resolution studies[3] and probably represent the true

© 2016 Macmillan Publishers Limited. All rights reserved

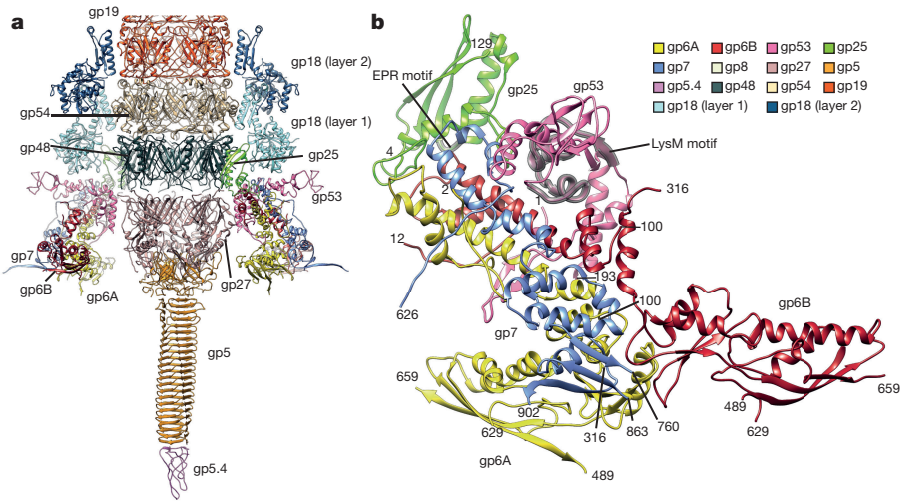


Figure 3 | Structure of the conserved inner baseplate. a, The minimal composition of a contractile injection system is derived from the T4 tail structure: the central spike complex (gp5, gp5.4, gp27), the conserved part of the wedge (gp6, gp7, gp25, gp53), the tail tube (gp19, gp48, gp54) and the conserved part of the sheath (gp18), some of which is modelled using the pyocin sheath[14]. **b**, A close-up of the structure of the conserved region of the wedge consisting of the $(gp6)_2$–gp7 heterotrimer, gp25 and gp53. The EPR motif of gp25 and the LysM motif of gp53 are highlighted with semi-transparent grey.

symmetry of the extended T4 tail. Gp54 and gp48 obey the helical symmetry of the tube. The rise and twist of superposition of the gp54 ring onto the gp19 and gp48 hexamers are 39.9 Å and 17.2°, and −38.8 Å and −22.1°, respectively.

In the complete tail, gp54 interacts with the first layer of sheath subunits (Fig. 3a), suggesting that gp54 is involved in the initiation of sheath assembly. However, most contractile injection systems lack gp54 orthologues and their first sheath layer must therefore bind to the main tube protein. Thus, the baseplate and the universally conserved gp25-like protein[1] in particular play a more important role in the initiation of sheath assembly in the extended state, as discussed further below.

Notably, T6SS gene clusters do not appear to contain orthologues of either gp48 or gp54 (Supplementary Table 3), presumably leaving both functions to Hcp, the main tube protein of T6SS. Hcp could acquire new functions by binding an adaptor or effector protein to its 'doughnut-hole' surface[40]. On the other hand, bacteria often carry several copies of *hcp* genes that can encode products with different functions.

The tail fibre network

The T4 baseplate distinguishes itself from its orthologues by being equipped with a very complex tail fibre network (Fig. 4). It consists of a gp9 trimer ($(gp9)_3$) to which a long tail fibre (LTF) comprising $(gp34)_3$,

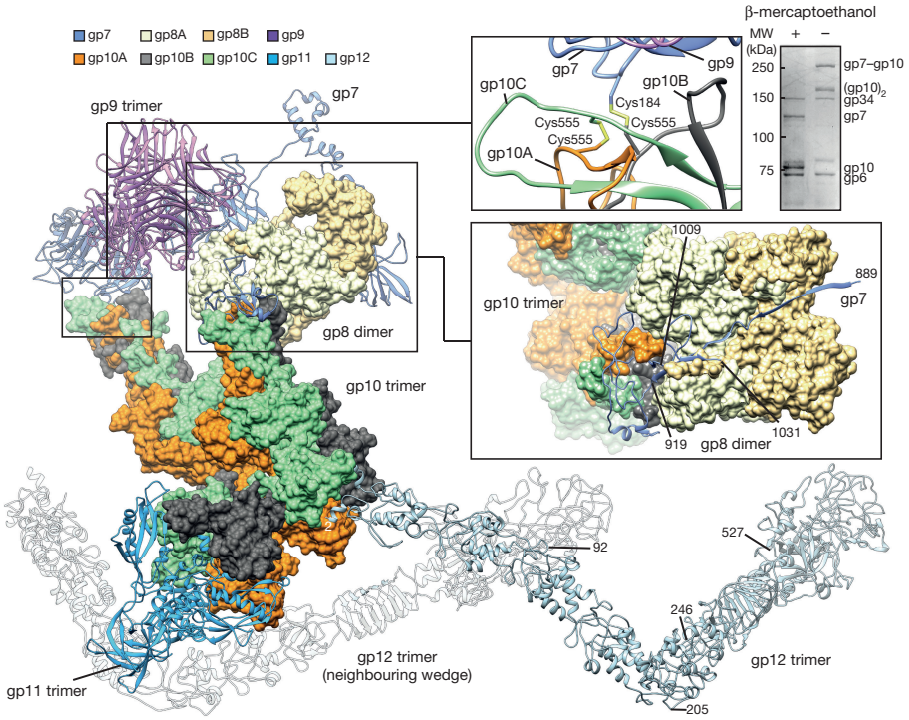


Figure 4 | The intermediate baseplate and the tail fibre network. Shown are gp7, gp8 (surface representation), gp9, gp10 (surface representation), gp11 and gp12. A symmetry-related gp12 trimer is displayed semi-transparently. Insets: 1) gp10–gp10 and gp10–gp7 disulfide bridges; 2) SDS–PAGE of T4 baseplates in reducing and non-reducing conditions (the full gel is shown in Supplementary Fig. 1); 3) close-up of the interaction of gp7 with gp8. Residues 890–918 of gp7 form a long β-strand that runs across the surface of the gp8 dimer end-to-end.

© 2016 Macmillan Publishers Limited. All rights reserved

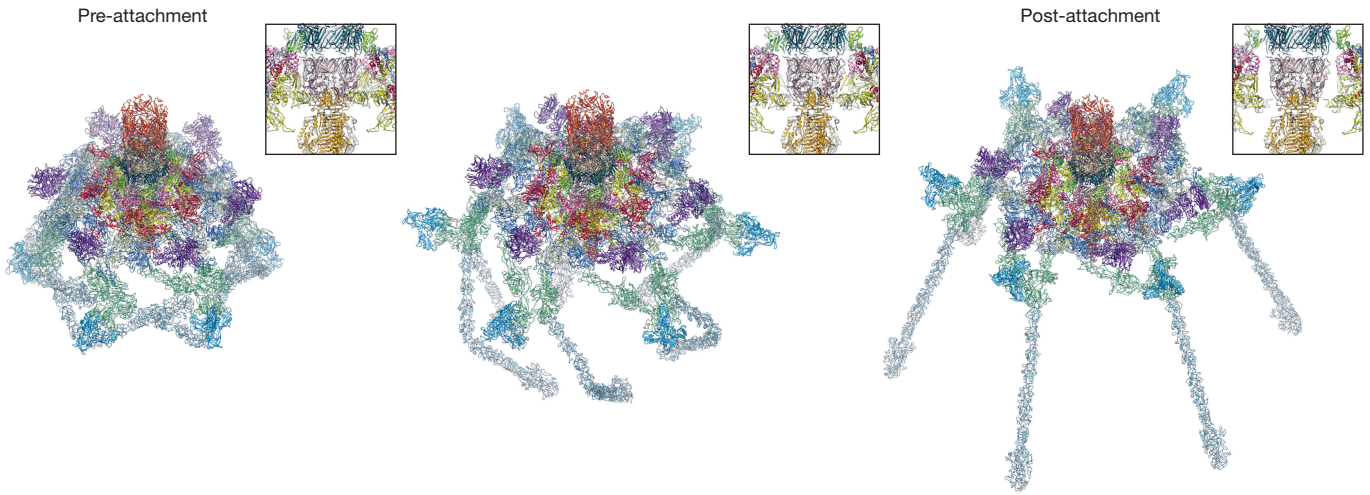


Figure 5 | Conformational change of the T4 baseplate upon host cell attachment. The left and right panels show the pre- and post-attachment structures that are derived from cryo-EM data. The middle image is a model of an intermediate with assumptions described in the text. The insets show close-up views of the central part of the baseplate demonstrating a release of the tail tube–central spike complex, whose position is kept unchanged throughout the transformation.

$(gp35)_1$, $(gp36)_3$ and $(gp37)_3$ is attached, and a $(gp10)_3$–$(gp11)_3$–$(gp12)_3$ module, in which $(gp12)_3$ is the short tail fibre (STF). The $(gp10)_3$–$(gp11)_3$–$(gp12)_3$ network forms a spring-like mechanism that extends upon interaction of the phage with a host cell. Our baseplate–tail tube complexes represent an intermediate of phage particle assembly and thus lack LTFs (except for a few aberrantly bound ones) because LTFs attach to otherwise complete phage particles at the very end of the particle assembly pathway[23]. The cryo-EM density of the LTF-attachment protein gp9 is weak but could nevertheless be easily interpreted with the help of its crystal structure[41] (Extended Data Fig. 2i).

Gp10 is key to understanding the function and evolution of the tail fibre network (Extended Data Fig. 4). This trimeric molecule has a distorted letter X shape, but each of its four domains (D1–D4) displays clear three-fold symmetry. A chain order swapping event takes place between D2 and D3 (Extended Data Fig. 4c). D1 is homologous to the N-terminal part of gp9 and this relationship could be detected at the level of the amino acid sequence. Unexpectedly, however, gp10 D2 and D3 are structurally similar to each other. They interact with the N-terminal motifs of gp12 and gp11, respectively, which themselves are also homologous (Extended Data Fig. 5a). Moreover, each of these units has a common three-fold axis (for example, gp10 D2–gp12) that orients the axes of the gp11 and gp12 trimers to be roughly perpendicular to each other. The X-like shape of gp10 allows each gp10–gp11 complex to interact with two STFs in the baseplate. One of the STFs extends D2 of gp10, and the other wraps around gp11 as described below.

Gp10 domains D1 and D4 attach the $(gp10)_3$–$(gp11)_3$–$(gp12)_3$ module to gp7 in the intermediate baseplate. Our structure shows, and mass spectrometry confirms, that apart from non-covalent interactions, the gp10 D4–gp7 interface involves a disulfide bridge between Cys555 of one gp10 chain and Cys184 of gp7 (Fig. 4). The Cys555 residues of the two remaining gp10 molecules form a disulfide bond (Extended Data Fig. 5b). This finding explains the gp10–gp10 inter-chain disulfide bridge found in the recombinant $(gp10)_3$–$(gp11)_3$ complex[42]. The gp7–gp10 and gp10–gp10 crosslinks are clearly seen in purified baseplates (Fig. 4). Cys555 of Gp10 and Cys184 of gp7 are conserved in all T4-like phages (Extended Data Fig. 5c, d), making the covalent attachment of the tail fibre network to the intermediate baseplate a hallmark of these viruses.

Using a multistep modelling procedure (see Methods), we were able to build the complete structure of the STF (Fig. 4 and Extended Data Fig. 6). Its previously unknown part (residues 2–245)[43,44] contains six repeats (Extended Data Fig. 6a–c) that follow the N-terminal gp11-like motif (residues 2–47; Extended Data Fig. 5a). Two kinks allow the fibre to fold neatly around the periphery of the baseplate. The first bend (approximately 30°) is between repeats 1 and 2 where the fibre contacts the C-terminal domain of a symmetry-related mate. The second, much sharper bend (approximately 90°) is between repeats 3 and 4 where the fibre wraps around the gp11 trimer. In this region, the gp11-proximal chain of gp12 folds into an α-helix whereas the other two chains adopt an extended conformation.

The structural data presented above show that all trimeric proteins comprising the baseplate periphery and fibres contain structurally related domains (Extended Data Fig. 6d). Thus, despite their different functions in the T4 particle, it is possible that all these proteins evolved from a common ancestor by multiple gene duplication events. Notably, simpler baseplates contain only one trimeric protein (the host receptor binding fibre) that is attached to a $(gp6)_2$–gp7-like unit[1].

The inner baseplate is universally conserved

The structure of the T4 baseplate allows us to define the composition of baseplates in other contractile injection systems. Using HMM profiles and structures of T4 baseplate proteins in HHsuite[45,46], we found orthologues of all inner baseplate components in all systems except for a gp53-like component in T6SS (Supplementary Table 3). This component could nevertheless constitute a domain in one of the T6SS core bundle proteins (TssF or TssG) as both are larger than some of their phage orthologues. Such a rearrangement of functional domains is not uncommon and, for example, gp102 of bacteriophage A511 combines the functions of gp25 and gp53 into a single protein (Supplementary Table 3).

To test the validity of our bioinformatics-based assignment of functions, we examined the assembly of the T6SS baseplate, as it appears to be more distantly related to T4 than other systems. We found that, similar to T4, the assembly pathway of the T6SS baseplate involves a wedge-like intermediate that contains a $(gp6)_2$–gp7-like heterotrimer (Extended Data Fig. 7a–d; for additional details regarding T4 baseplate assembly, see Supplementary Information). The T6SS orthologues of T4 gp25, gp6 and gp7 (TssE, TssF and TssG, respectively) form a complex with TssK that has a stoichiometry of $(TssE)_1$–$(TssF)_2$–$(TssG)_1$–$(TssK)_3$ (Extended Data Fig. 7e, f). No complex was detected in the absence of TssK.

Considering the strength of bioinformatic predictions and the biochemical data, we conclude that the TssE–$(TssF)_2$–TssG unit forms the core bundle, the trifurcation unit, and the sheath polymerization platform in the T6SS baseplate. Furthermore, similar to the function of the gp10 trimer in T4 (ref. 23), trimeric TssK (ref. 47) appears to initiate the assembly of the T6SS wedge.

© 2016 Macmillan Publishers Limited. All rights reserved

Conformational switch of the baseplate

Although the structure of a single state of a macromolecular machine as complex as the T4 baseplate can provide a plethora of information, to understand fully how it works one needs to describe the machine in its multiple states. Fortuitously, a small fraction of the baseplates in our study switched to the post-host attachment state (Extended Data Fig. 1). We selected all such particles and calculated a reconstruction to 6.8 Å resolution. Virtually all domains from the pre-attachment hexagonal baseplate fit into this map as rigid bodies (see Methods). One notable exception is a loop formed by residues 841–862 of gp7. Cryo-EM density convincingly places this loop on two opposite sides of the gp7 trifurcation unit insertion domain in the two conformations (Extended Data Fig. 8a, b), indicating that it performs a 'jump rope' type of motion around this domain during the conformational switch. The location of this loop in our post-attachment structure is similar to that found in the 3.8 Å resolution reconstruction of the recombinantly assembled star-shaped baseplate[48].

To better understand and visualize structural changes that relate the two conformations, we aligned the two structures using gp53 molecules as a reference (see Methods) and performed interpolation with the help of UCSF Chimera[49] (Fig. 5 and Supplementary Video 1). This procedure resulted in intermediates with no grave clashes or chain breaks although it did not work for the STFs or the jump rope loop of gp7. Several intermediates had to be created as described in Methods.

All baseplate proteins participate in the conformational switch (Fig. 5 and Supplementary Video 1). The tail fibre network displays the largest amount of motion. The STFs unravel and straighten, as can be seen in raw images of baseplates attached to membranes (Extended Data Fig. 2j). The multivalent structure of the gp10 trimer plays a critical role in this transformation. Three-fold symmetrical gp10–gp11 and gp10–gp12 interfaces are preserved in the two conformations, suggesting that the gp10–gp11–gp12 complex rotates as a unit during the transformation. At the end of this rotation, the N-terminal domain of gp12 points straight towards the host cell surface. On the other hand, many interactions that lack three-fold symmetry in the pre-attachment state are broken: gp11 dissociates from the gp12 fibre that was originally wrapped around it (gp11 then binds the LTF protein gp34 in the post-attachment baseplate[4]); and the C-terminal domain gp12 dissociates from a symmetry-related gp12 fibre and binds to a receptor on the host cell surface.

The reorganization of the inner baseplate can be best understood by examining the motion of $(gp6)_2$–gp7 heterotrimers (Extended Data Fig. 9 and Supplementary Videos 2, 3). Gp7 transfers the rotation of the gp10–gp11–gp12 module to the $(gp6)_2$–gp7 heterotrimer, causing the latter also to rotate as a unit and tilt its three-fold axis by about 30°. As a result, the diameter of the $(gp6)_2$–gp7 iris increases and the proteins bound to the core bundle (gp25 and gp53) dissociate from the tube and the hub (Fig. 5), preparing the latter for subsequent sheath-driven translocation.

To analyse how the relative positions of different baseplate domains influence the baseplate conformational change and phage infectivity, we subjected gp7 to site-directed mutagenesis by extending or shortening some of its domain-linking regions by one amino acid (Extended Data Fig. 8 and Supplementary Information). We also attempted to remove the gp7–gp10 disulfide bridge. The region preceding the insertion domain of the trifurcation unit did not tolerate the two desired mutations, whereas small changes in the length of the linker upstream of the trifurcation unit and the jump rope loop did not affect phage infectivity (Extended Data Fig. 8). On the other hand, removal of the gp7–gp10 covalent linkage had a profound effect on infectivity, confirming that this bond is important for maintaining the structure of the baseplate during the conformational switch.

To extrapolate our results to an even larger framework, we used all available structural and functional data[19,50] to construct an atomic model of the T4 tail in the extended and contracted states (see Methods, Extended Data Fig. 10). Gp25, gp54 and the gp6B–gp53 platform create a complex interface that promotes the assembly of the first layer of the sheath in the extended conformation (Extended Data Fig. 10a). Gp25 serves as a nucleus for sheath polymerization because the structure and symmetry of the gp25–gp48 complex are repeated in every sheath–tube subunit interaction. In the assembled structure, gp25 is a seamless continuation of the lattice formed by gp25-like tube-interacting domains of the sheath. To achieve this, gp25 must accept the long N- and C-terminal arms of the first sheath subunit[14,31,32], thus firmly attaching the sheath to the baseplate (Extended Data Fig. 10; for additional details see Supplementary Information). In the contracted tail and the post-attachment baseplate, an outward motion and rotation of the core bundle with the associated proteins (gp25 and gp53) causes an even larger outward motion and around 45° rotation of sheath subunits[27] (Extended Data Fig. 10 and Supplementary Video 2). The sheath appears to maintain its original interaction with gp25, but now finds a different, presumably more energetically favourable, interface on the gp6B–gp53 platform (Extended Data Fig. 10b). Thus, in agreement with its high conservation in contractile injection systems[1], gp25 plays a critical role in sheath assembly and contraction.

The structural transformation of the baseplate described here involves proteins that are universally conserved in all contractile injection systems (except for the unknown T6SS gp53 orthologue). A similar sequence of events is likely to occur in any such system, regardless of the complexity of its peripheral baseplate or tail fibre network. Thus, the conformational switch consists of the following steps: an external factor triggers reorganization of the peripheral baseplate; the gp7 orthologue transfers this signal to a gp25– $(gp6)_2$–gp7-like module; and this module reorients, releasing the hub complex while simultaneously initiating sheath contraction.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 3 February; accepted 5 April 2016.

1. Leiman, P. G. & Shneider, M. M. Contractile tail machines of bacteriophages. *Adv. Exp. Med. Biol.* **726,** 93–114 (2012).
2. Hu, B., Margolin, W., Molineux, I. J. & Liu, J. Structural remodeling of bacteriophage T4 and host membranes during infection initiation. *Proc. Natl Acad. Sci. USA* **112,** E4919–E4928 (2015).
3. Kostyuchenko, V. A. *et al.* The tail structure of bacteriophage T4 and its mechanism of contraction. *Nature Struct. Mol. Biol.* **12,** 810–813 (2005).
4. Leiman, P. G., Chipman, P. R., Kostyuchenko, V. A., Mesyanzhinov, V. V. & Rossmann, M. G. Three-dimensional rearrangement of proteins in the tail of bacteriophage T4 on infection of its host. *Cell* **118,** 419–429 (2004).
5. Kanamaru, S. *et al.* Structure of the cell-puncturing device of bacteriophage T4. *Nature* **415,** 553–557 (2002).
6. Browning, C., Shneider, M. M., Bowman, V. D., Schwarzer, D. & Leiman, P. G. Phage pierces the host cell membrane with the iron-loaded spike. *Structure* **20,** 326–339 (2012).
7. Leiman, P. G. *et al.* Type VI secretion apparatus and phage tail-associated protein complexes share a common evolutionary origin. *Proc. Natl Acad. Sci. USA* **106,** 4154–4159 (2009).
8. Basler, M. Type VI secretion system: secretion by a contractile nanomachine. *Phil. Trans. R. Soc. Lond. B* **370,** 20150021 (2015).
9. Shneider, M. M. *et al.* PAAR-repeat proteins sharpen and diversify the type VI secretion system spike. *Nature* **500,** 350–353 (2013).
10. Basler, M., Pilhofer, M., Henderson, G. P., Jensen, G. J. & Mekalanos, J. J. Type VI secretion requires a dynamic contractile phage tail-like structure. *Nature* **483,** 182–186 (2012).
11. Mougous, J. D. *et al.* A virulence locus of *Pseudomonas aeruginosa* encodes a protein secretion apparatus. *Science* **312,** 1526–1530 (2006).
12. Shikuma, N. J. *et al.* Marine tubeworm metamorphosis induced by arrays of bacterial phage tail-like structures. *Science* **343,** 529–533 (2014).
13. Nakayama, K. *et al.* The R-type pyocin of *Pseudomonas aeruginosa* is related to P2 phage, and the F-type is related to lambda phage. *Mol. Microbiol.* **38,** 213–231 (2000).
14. Ge, P. *et al.* Atomic structures of a bactericidal contractile nanotube in its pre- and postcontraction states. *Nature Struct. Mol. Biol.* **22,** 377–382 (2015).
15. Heymann, J. B. *et al.* Three-dimensional structure of the toxin-delivery particle antifeeding prophage of *Serratia entomophila*. *J. Biol. Chem.* **288,** 25276–25284 (2013).
16. Yang, G., Dowling, A. J. & Gerike, U. ffrench-Constant, R. H. & Waterfield, N. R. Photorhabdus virulence cassettes confer injectable insecticidal activity against the wax moth. *J. Bacteriol.* **188,** 2254–2261 (2006).

© 2016 Macmillan Publishers Limited. All rights reserved

17. Yamamoto, T. Presence of rhapidosomes in various species of bacteria and their morphological characteristics. *J. Bacteriol.* **94,** 1746–1756 (1967).
18. Bönemann, G., Pietrosiuk, A. & Mogk, A. Tubules and donuts: a type VI secretion story. *Mol. Microbiol.* **76,** 815–821 (2010).
19. Leiman, P. G. *et al.* Morphogenesis of the T4 tail and tail fibers. *Virol. J.* **7,** 355 (2010).
20. Kikuchi, Y. & King, J. Genetic control of bacteriophage T4 baseplate morphogenesis. I. Sequential assembly of the major precursor, *in vivo* and *in vitro*. *J. Mol. Biol.* **99,** 645–672 (1975).
21. Kikuchi, Y. & King, J. Genetic control of bacteriophage T4 baseplate morphogenesis. II. Mutants unable to form the central part of the baseplate. *J. Mol. Biol.* **99,** 673–694 (1975).
22. Kikuchi, Y. & King, J. Genetic control of bacteriophage T4 baseplate morphogenesis. III. Formation of the central plug and overall assembly pathway. *J. Mol. Biol.* **99,** 695–716 (1975).
23. Coombs, D. H. & Arisaka, F. in *Molecular Biology of Bacteriophage T4* (ed. Karam, J. D.) 259–281 (American Society for Microbiology, 1994).
24. Kikuchi, Y. & King, J. Assembly of the tail of bacteriophage T4. *J. Supramol. Struct.* **3,** 24–38 (1975).
25. De Rosier, D. J. & Klug, A. Reconstruction of three dimensional structures from electron micrographs. *Nature* **217,** 130–134 (1968).
26. Kostyuchenko, V. A. *et al.* Three-dimensional structure of bacteriophage T4 baseplate. *Nature Struct. Biol.* **10,** 688–693 (2003).
27. Aksyuk, A. A. *et al.* The tail sheath structure of bacteriophage T4: a molecular machine for infecting bacteria. *EMBO J.* **28,** 821–829 (2009).
28. Amos, L. A. & Klug, A. Three-dimensional image reconstructions of the contractile tail of T4 bacteriophage. *J. Mol. Biol.* **99,** 51–64 (1975).
29. Effantin, G. *et al.* Cryo-electron microscopy three-dimensional structure of the jumbo phage PhiRSL1 infecting the phytopathogen *Ralstonia solanacearum*. *Structure* **21,** 298–305 (2013).
30. Liu, J., Chen, C. Y., Shiomi, D., Niki, H. & Margolin, W. Visualization of bacteriophage P1 infection by cryo-electron tomography of tiny *Escherichia coli*. *Virology* **417,** 304–311 (2011).
31. Kudryashev, M. *et al.* Structure of the type VI secretion system contractile sheath. *Cell* **160,** 952–962 (2015).
32. Clemens, D. L., Ge, P., Lee, B. Y., Horwitz, M. A. & Zhou, Z. H. Atomic structure of T6SS reveals interlaced array essential to function. *Cell* **160,** 940–951 (2015).
33. Crowther, R. A., Lenk, E. V., Kikuchi, Y. & King, J. Molecular reorganization in the hexagon to star transition of the baseplate of bacteriophage T4. *J. Mol. Biol.* **116,** 489–523 (1977).
34. Amunts, A., Brown, A., Toots, J., Scheres, S. H. & Ramakrishnan, V. Ribosome. The structure of the human mitochondrial ribosome. *Science* **348,** 95–98 (2015).
35. Yan, C. *et al.* Structure of a yeast spliceosome at 3.6-angstrom resolution. *Science* **349,** 1182–1191 (2015).
36. Aksyuk, A. A., Leiman, P. G., Shneider, M. M., Mesyanzhinov, V. V. & Rossmann, M. G. The structure of gene product 6 of bacteriophage T4, the hinge-pin of the baseplate. *Structure* **17,** 800–808 (2009).
37. Maxwell, K. L. *et al.* Structural and functional studies of gpX of *Escherichia coli* phage P2 reveal a widespread role for LysM domains in the baseplates of contractile-tailed phages. *J. Bacteriol.* **195,** 5461–5468 (2013); correction **196,** 2122 (2014).
38. Lossi, N. S., Dajani, R., Freemont, P. & Filloux, A. Structure-function analysis of HsiF, a gp25-like component of the type VI secretion system, in *Pseudomonas aeruginosa*. *Microbiology* **157,** 3292–3305 (2011).
39. Veesler, D. & Johnson, J. E. Virus maturation. *Annu. Rev. Biophys.* **41,** 473–496 (2012).
40. Silverman, J. M. *et al.* Haemolysin coregulated protein is an exported receptor and chaperone of type VI secretion substrates. *Mol. Cell* **51,** 584–593 (2013).
41. Kostyuchenko, V. A. *et al.* The structure of bacteriophage T4 gene product 9: the trigger for tail contraction. *Structure* **7,** 1213–1222 (1999).
42. Zhao, L., Takeda, S., Leiman, P. G. & Arisaka, F. Stoichiometry and inter-subunit interaction of the wedge initiation complex, gp10-gp11, of bacteriophage T4. *Biochim. Biophys. Acta* **1479,** 286–292 (2000).
43. Thomassen, E. *et al.* The structure of the receptor-binding domain of the bacteriophage T4 short tail fibre reveals a knitted trimeric metal-binding fold. *J. Mol. Biol.* **331,** 361–373 (2003).
44. van Raaij, M. J., Schoehn, G., Burda, M. R. & Miller, S. Crystal structure of a heat and protease-stable part of the bacteriophage T4 short tail fibre. *J. Mol. Biol.* **314,** 1137–1146 (2001).
45. Söding, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21,** 951–960 (2005).
46. Remmert, M., Biegert, A., Hauser, A. & Soding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods* **9,** 173–175 (2012).
47. Zoued, A. *et al.* TssK is a trimeric cytoplasmic protein interacting with components of both phage-like and membrane anchoring complexes of the type VI secretion system. *J. Biol. Chem.* **288,** 27031–27041 (2013).
48. Yap, M. L. *et al.* Role of bacteriophage T4 baseplate in regulating assembly and infection. *Proc. Natl Acad. Sci. USA* **113,** 2654–2659 (2016).
49. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).
50. Fokine, A. *et al.* The molecular architecture of the bacteriophage T4 neck. *J. Mol. Biol.* **425,** 1731–1744 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank V. Mesyanshinov for developing the initial baseplate–tail tube complex purification protocol; C. Maillard for technical support; A. Brown for advice on structure refinement involving cryo-EM data; D. Demurtas for sample screening by negative stain EM; R. McLeod for help with data transfer; S. Nazarov, M. Plattner, and V. Kostyuchenko for discussions; and M. Basler for reading the manuscript. We acknowledge support from the EPFL SCITAS (high performance computing), the EPFL Centre for Interdisciplinary Electron Microscopy and the EPFL Proteomics Core Facility. The work was supported by the Swiss National Science Foundation grant 310030_144243.

Author Contributions N.M.I.T. purified the baseplates, calculated cryo-EM reconstructions, built and refined all atomic models, performed bioinformatic analyses and wrote the first draft of the paper. R.C.G.-F., N.M.I.T., K.N.G. and H.S. collected cryo-EM data. N.M.I.T. and N.S.P. designed, and N.S.P. performed, site-directed mutagenesis of T4 phage. M.M.S. designed, produced and analysed T6SS samples and T4 gp25. C.B. crystallized T4 gp25. P.G.L. solved the T4 gp25 structure, analysed the data from all sources, and integrated all the information into a single manuscript.

Author Information Cryo-EM maps have been deposited in the Electron Microscopy Data Bank under the following accession numbers: EMD-3374 and EMD-3396 for the pre-attachment and post-attachment baseplate, respectively, and EMD-3397, EMD-3392, EMD-3393, EMD-3394 and EMD-3395 for the locally masked reconstructions of the pre-attachment baseplate for the tail tube, inner, intermediate, upper and lower peripheral baseplate regions, respectively. Atomic coordinates have been deposited in the Protein Data Bank under the following accession numbers: 5IV5 and 5IV7 for the pre-attachment and post-attachment baseplate, respectively, and 4HRZ for the crystal structure of T4 gp25. The vector with the TEV-cleavable His–SlyD expression tag derived from pET-23d(+) (Novagen) has been deposited to the NCBI database under the accession number KU314761. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.G.L. (petr.leiman@epfl.ch).

© 2016 Macmillan Publishers Limited. All rights reserved

METHODS

Purification of the T4 baseplate–tail tube complex. We purified T4 baseplates by improving a protocol established in a previous study[26] by implementing new tools, resins and media (Extended Data Fig. 1a). This protocol involves a phage mutant that carries the amber stop codon in the 5′ region of the sheath and major capsid protein genes (T4-18am-23am). T4 particle assembly is ordered and infection of a non-amber-suppressing (Sup^0) cell strain by such a mutant results in accumulation of assembly intermediates up to the point of incorporation of the defective gene product (that is, baseplate–tail tube complexes and fibres in the case of T4-18am-23am; Extended Data Fig. 1b, c).

Six litres of Sup^0 *Escherichia coli* B^E cells were grown at 37 °C to an attenuance of 0.6 at 600 nm ($A_{600} = 0.6$) and infected by T4-18am-23am with a multiplicity of infection (M.O.I. or phage-to-cell ratio) of 5. Eight minutes later, the same amount of phage particles were added, resulting in lysis delay[51] or 'superinfection'. The infected cells were incubated at 30 °C for an additional 150 min, collected by centrifugation and flash frozen in liquid nitrogen. The cells were then thawed and resuspended in 120 ml of lysis buffer (50 mM Tris-HCl pH 8.0, 1 mM $MgCl_2$, 30 μg ml^{-1} DNase I, 100 μg ml^{-1} lysozyme), and incubated at 30 °C for 30 min. Then, an equal volume of RNase buffer (50 mM Tris-HCl pH 8.0, 40 mM EDTA) containing RNase at 50 μg ml^{-1} was added, and the cells were incubated at 30 °C for an additional 30 min.

This sample was concentrated to ~20 ml with the help of Centricon Plus-70 (30 kDa cut-off) filtration devices (Merck Millipore) and centrifuged at 50,000*g* overnight. The pellet was resuspended in purification buffer (50 mM Tris-HCl pH 8.0, 1 mM $MgCl_2$), loaded on a sucrose gradient (5–30% (w/v) with a 50% cushion in the same buffer), and centrifuged at 35,000*g* for 4 h. Fractions containing baseplates were pooled, diluted with the same buffer and centrifuged at 50,000*g* overnight. The pellet was resuspended in SM8 buffer (50 mM Tris-HCl pH 8.0, 100 mM NaCl, 8 mM $MgSO_4$), loaded on an OptiPrep (PROGEN Biotechnik) gradient (10–40% (w/v) with a 50% cushion) buffered by SM8 buffer and centrifuged at 35,000*g* for 24 h. Fractions containing baseplates were pooled and subjected to size-exclusion chromatography to remove the OptiPrep compound. A Superose 6 Increase 10/300 GL column (GE Healthcare Life Sciences) pre-equilibrated with SM8 buffer was used. Peak fractions were pooled and centrifuged at 50,000*g* for 4 h. The pellet was resuspended in about 200 μl of SM8 buffer, to a final concentration of around 1 mg ml^{-1}.

Cryo-EM sample preparation and data acquisition. T4 baseplates were applied to glow-discharged, 300-mesh, copper Quantifoil grids (Quantifoil). The grids were blotted for 3 s and vitrified by plunging into liquid ethane cooled down to liquid nitrogen temperature using a Vitrobot Mark IV (FEI). Cryo-EM samples were imaged using a K2 summit direct electron detector (Gatan) on an FEI Titan Krios at an acceleration voltage of 300 kV and equipped with a Gatan image filter (slit width 20 eV) (Supplementary Table 2). Images were collected at a defocus of around 1.5 μm and recorded in dose fractionation mode (40 frames) at a dose rate between 6 and 8 e^- per pixel per s, dose per image sub-frame between 1 and 1.5 e^- $Å^{-2}$ and a cumulative dose between 40 and 60 e^- $Å^{-2}$. The effective pixel size of the images was 1.32 Å per pixel at a nominal magnification of 105,000×. Frames 3–40 were aligned by motion correction[52] as implemented in 2dx_automator[53].

Cryo-EM data processing. For all 1,699 micrographs, we calculated the defocus parameters using CTFFind4 (ref. 54) and discarded 78 micrographs that fell outside a 0.5–4-μM defocus window, had a cross-correlation of lower than 0.006 or did not have Thon rings extending to at least 8 Å. 47,516 pre-attachment and 6,900 post-attachment particles were picked using e2boxer.py in EMAN2 (ref. 55).

For the pre-attachment baseplate, particles were classified into 250 classes using reference-free 2D classification with RELION 1.4 (ref. 56). The 47,362 particles belonging to the best 41 classes were regrouped into 421 groups, and used to calculate the initial reconstruction using RELION 3D auto-refine. The previously published 12 Å resolution baseplate–tail tube complex (EMDB code 1048)[26] was filtered to 60 Å and used as an initial reference. This procedure gave a reconstruction with a resolution of 5.35 Å before and 4.36 Å after postprocessing employing automatic masking, with an empirically optimized binarization threshold of 0.11, automated B-factor sharpening and correction of the modulation transfer function (MTF)[57]. To further improve the resolution, we performed alignment of the individual particle movie frames ('particle polishing'), using 47,297 particles after discarding those that had missing frames. This was followed by a new refinement using these polished ('shiny') particles. We regrouped the particles into 19 groups and repeated the 3D refinement. Then, particles with a maximum probability distribution value lower than 0.09 were discarded and the final round of refinement was run using 37,913 particles. This resulted in a reconstruction with a resolution of 4.65 Å. This map was postprocessed using automatic masking, an empirically optimized binarization threshold of 0.11, automated B-factor sharpening and correction of the modulation transfer function (MTF). The postprocessed map had an overall resolution of 4.11 Å (ref. 57).

To perform localized reconstructions, we low-pass filtered the map from the final round of refinement to 8 Å, generated local masks by segmentation with Segger[58] in UCSF Chimera[49], and processed them with relion_mask_create. The subsequent refinement was performed by applying these local masks[34], similar to Yan *et al.*[35]. This procedure resulted in maps with generally improved interpretability for most regions of the baseplate, with the exception of the central non-six-fold-symmetric central spike and the gp9 protein, which displays a very high conformational heterogeneity. These maps after postprocessing, and their FSC curves, are shown in Extended Data Fig. 2g, h.

Reconstruction of the post-attachment baseplate was initiated by grouping all images into 70 classes with RELION while limiting the high-resolution terms to 10 Å and performing phase flipping instead of the full CTF correction. The 5,188 particles belonging to the best 25 classes were selected and regrouped into 25 groups. These particles were then used for RELION 3D refinement. We created the starting model by extracting the baseplate from the previously published 17 Å reconstruction of the contracted T4 phage tail (EMDB code 1089)[4] with the help of Segger[58] in UCSF Chimera[49]. This map was filtered to 60 Å and subjected to the RELION 3D auto-refine procedure, resulting in a 9.50 Å resolution reconstruction. Postprocessing with an optimized binarization threshold of 0.15 improved the resolution to 7.49 Å. A total of 5,176 particles had complete stacks of movie frames, which allowed us to 'polish' them by aligning individual movie frames. The resulting map had a resolution of 6.77 Å. We further improved the interpretability of this map by postprocessing using the mask from the polishing procedure, MTF correction and automatic B-factor correction.

Model building: pre-attachment baseplate. All available crystal structures of baseplate proteins and their domains[5,36,41,43,44,59–61] were fitted into the best postprocessed map using Coot[62] (Supplementary Table 1). Reliable homology models were generated with the help of HHpred[63] and Phyre2[64], fitted into the map with Coot and rebuilt. The rest of the models were built *de novo* also in Coot. Ramachandran restraints were used for local real space refinement, which was performed against the postprocessed overall map or the postprocessed localized refinement maps (which were aligned to the postprocessed overall map, and resampled using Chimera). For the tail tube protein gp19, we generated a model using MODELLER[65] based on HHalign[45] alignment with gp54. Finally, taking advantage of the technological development of the last 15 years[66], we re-refined the published structure of the gp12 fragment[44]. The improved electron density allowed us to extend the existing model and complete the structure of the gp12 repeat unit (residues 220–253; Extended Data Fig. 6b, c). This structure was then used to model the other repeats with MODELLER, imposing three-fold symmetry restraints. The N-terminal residues of gp12 were modelled on the basis of alignment with the gp11 N-terminal residues found by HHpred, using MODELLER and three-fold symmetry restraints.

The cryo-EM map of gp10 domains 2 and 3 was one of the most challenging regions for model building. Therefore, we used mass spectroscopy to verify the *de novo* built model. A disulfide bond between cysteines 331 and 342 was first identified in our model and then confirmed by mass spectrometry, indicating the correctness of the atomic model even in difficult parts of electron density.

Model building: post-attachment baseplate. The model of the post-attachment baseplate was constructed with the help of the refined coordinates of the pre-attachment baseplate in the following way. Initially, the inner baseplate (gp6 (residues 2–104, 170–278 and 317–411), gp7 (residues 638–753), gp25 and gp53) was fitted as a rigid body. Then, gp8 and gp9 were located and the remaining domains of gp6 and gp7 were fitted with the help of the secondary structure elements visible in the map. Residues 841–862 of gp7 (the jump rope loop) had to be completely rebuilt.

In the peripheral baseplate, the following modules were fitted as rigid units: gp10 domain 1, (gp10 domain 2)–(gp12 N-terminal domain), (gp10 domain 3)–gp11, and gp10 domain 4. Subsequently, linker regions and a few loops were remodelled by rigid body fitting, regularization and real space refinement.

The final refined model does not contain gp12. The structure of the extended gp12 STF is obtained by assuming that the interaction between gp10 domain 2 and the N-terminal domain of gp12 is preserved and that gp12 simply straightens upon interaction with the host cell membrane (Extended Data Fig. 2j).

Refinement. Both structures (pre- and post-attachment baseplate) were refined with phenix.real_space_refine[67] using global real space refinement and reciprocal space B-factor refinement (two B-factors per residue; Supplementary Table 2). The following restrains were applied in the refinement: standard C_β, secondary structure, Ramachandran, global six-fold symmetry and local two- and three-fold symmetry for the following proteins: gp9, gp11, and gp19. In the pre-attachment structure, the central spike $(gp5.4)_1$–$(gp5)_3$–$(gp27)_3$ was modelled as alternate conformers in its six different symmetry-related positions. The pre-attachment

© 2016 Macmillan Publishers Limited. All rights reserved

structure served as a reference model in the refinement of the post-attachment structure providing additional restraints.

To detect overfitting, we performed a limited random displacement of atoms in the final model with a mean of 0.5 Å and refined it against one of the two maps from the final round of cryo-EM image processing, each calculated with half of the data. Both models showed similar agreement with both maps confirming that no overfitting took place[68] (Extended Data Fig. 2c, f).

Baseplate conformational change modelling. The post-attachment baseplate was aligned to the pre-attachment baseplate by least-squares superposition of the centres of mass of six gp53 molecules in the two conformations using LSQKAB[69] from the *CCP*4 suite of programs[70]. Then, we calculated a morph between the pre-and post-attachment baseplate by means of the corkscrew interpolation algorithm as implemented in UCSF Chimera[49]. Only two regions of the entire baseplate structure had physically impossible intermediates: the three chains comprising the STF had independent trajectories resulting in their spatial separation, and the jump rope loop of gp7 (residues 841–862) overlapped with other atoms of the gp7 trifurcation unit insertion domain (residues 761–862). Splitting the STF into two separate trimeric parts comprising residues 2–98 and 99–527 resulted in a trajectory in which all three chains maintained their native association, but the fibre was now separated at its sharp kink region (residues 180–185). To overcome this problem, we created intermediates in which the kink region was manually straightened and regularized. Finally, the two ends of the jump rope loop of gp7 were rebuilt to allow its large motion around the trifurcation unit insertion domain, and the insertion domain had to be slightly displaced as a rigid body to accommodate this movement.

Pseudoatomic tail models: extended tail. The pseudo-atomic model of the T4 tail presented in Extended Data Fig. 10 was produced by fitting the structures—as rigid bodies—of tail proteins obtained in this and other studies[5,19,27,36,41,43,44,50,59–61,71] into the cryo-EM map of the extended T4 tail (EMBD code 1126)[3] with the help of UCSF Chimera[49]. A crystal structure of the sheath protein gp18 lacking gp25-like Domain I[1] and a part of Domain II[1] is available (PDB code 3FOA)[27]. We used HHpred[63], MODELLER[65], and the structure of the R-type pyocin sheath (PDB code 3J9Q)[14] to model the missing fragments giving rise to a nearly complete atomic model of the sheath protein subunit. This structure was then fitted into the tail cryo-EM map. The tail tube was constructed by repetitively applying the symmetry transformation relating the first two gp19 hexamers in the atomic structure described in this work. The atomic model of the tail tube terminator protein gp3 (ref. 72) was derived from the structure of gp19 using the same tools[63,65] and was assumed to obey the overall symmetry of the tube.

Pseudoatomic tail models: contracted tail. Several cryo-EM reconstructions of the T4 tail in the contracted state are available (EMDB codes 1086, 1089, 5528)[4,50]. Of these, the highest resolution (EMDB code 1086) is at 17 Å resolution[4], but it lacks the fibritin whisker proteins, which are present in a more recent study that has a resolution of 25 Å (EMDB code 5528)[50]. To facilitate the interpretation of these complex maps, they were brought to a common origin with the structure of the post-attachment baseplate (which was itself aligned to the pre-attachment baseplate, as described above). We then fitted the atomic structure of the post-attachment baseplate, all the sheath subunits, and the tube from the extended tail into the 17 Å resolution map. The tail- and tube-terminating proteins (gp15 and gp3, respectively), were assumed to preserve their relative orientation in the contracted tail, thus defining the position of the tube. The fibritin model was then taken from the 25 Å resolution study.

Purification, crystallization and structure determination of T4 gp25. Our original attempt to express gp25 resulted in an insoluble protein (Extended Data Fig. 7g, lanes 2 and 3). To produce gp25 in soluble form, we expressed it as a downstream fusion to SlyD, a known metal-affinity column contaminant, peptidyl-prolyl isomerase and expression enhancer[73] (Extended Data Fig. 7g, lanes 4–8). The SlyD tag carried a polyhistidine sequence at its N terminus to increase its affinity to IMAC resin. The linker between SlyD and gp25 contained a cleavage site for the tobacco etch virus (TEV) protease. We used the TEV protease with a polyhistidine tag at its C terminus that we produce recombinantly in-house (Extended Data Fig. 7g, lane 9). The vector with the TEV-cleavable His–SlyD expression tag was derived from pET-23d(+) (Novagen). We named it pTSL and deposited it to the NCBI database under accession number KU314761.

For cloning into pTSL, gene *25* was PCR amplified using the primers given in Supplementary Table 4 and T4 DNA as a template. The expression vector was transformed into methionine-auxotrophic B834/DE3 cells with ampicillin as a selection marker. Protein expression was performed in 2 × TY medium supplemented with ampicillin at 200 mg ml^{-1}. The plasmid-carrying cells were grown to A_{600} of 0.6 and induced by addition of IPTG to a final concentration of 0.5 mM (Extended Data Fig. 7g, lanes 4, 5). The expression was performed at 18 °C with rigorous shaking overnight. The cells were pelleted at 8,000*g*, resuspended in lysis buffer (50 mM Tris-HCl pH 8.0, 300 mM NaCl), and disrupted by sonication at 4–16 °C. The insoluble fraction was removed by centrifugation at 27,000*g* for 20 min. The clarified cell lysate was loaded onto a 5-ml GE HisTrap FF Ni-charged column, and the protein was eluted with imidazole-containing buffer (50 mM Tris-HCl pH 8.0, 300 mM NaCl, 200 mM imidazole) using a step gradient (Extended Data Fig. 7g, lane 6). Fractions containing the protein were pooled and the TEV protease was added to give a protease/protein ratio of 1/100 (w/w). This mixture was dialysed into 10 mM Tris-HCl pH 8.0 by overnight incubation at 20 °C resulting in cleavage of the His–SlyD expression tag (Extended Data Fig. 7g, lane 7). The digested protein was filtered and purified by ion-exchange chromatography using a MonoQ 10/100 GL column and 0 to 1 M NaCl gradient in 20 mM Tris–HCl pH 8.0 buffer. Fractions containing gp25 were pooled and loaded onto a Superdex 75 HiLoad 16/60 size-exclusion column equilibrated with 10 mM Tris–HCl pH 8.0 and 150 mM NaCl.

For crystallization, the protein was dialysed into crystallization buffer (20 mM Tris–HCl at pH 8.0, 50 mM NaCl), concentrated to 20 mg ml^{-1}, and crystallized in 0.9–1.0 M $MgSO_4$, 100 mM MES pH 6.2. The structure was solved by the single-wavelength anomalous dispersion method with the help of a SelenoMethionine (SeMet) derivative. The latter was produced using the SelenoMethionine Medium Complete kit (Molecular Dimensions). The SeMet derivative was purified and crystallized in a similar way to the native protein. Crystallographic data were collected at the Swiss Light Source PX-I (X06SA) beamline using a PILATUS-6 M (DECTRIS Ltd.) pixel detector. For data collection, a crystal was briefly washed in cryoprotectant solution that contained all the components of the well solution at the same concentrations and 25% ethylene glycol and flash frozen in a vaporized liquid N_2 stream at 100 K. The X-ray data were processed with MOSFLM[74], SCALA[75] and POINTLESS[76] (Supplementary Table 5). The selenium atoms were located with the help of SHELXD[77]. The atomic model was built and refined with COOT[62] and PHENIX[78].

Purification of TssE–TssF–TssG–TssK complex. Recombinant expression of *E. coli* CFT073 gene c3400, which was proposed to encode a TssF–TssG fusion[79], results in two proteins with molecular masses of ∼65 kDa and ∼40 kDa. The sequence of CFT073 'gene' *c3400* is identical to that of genes *CP009072* of *E. coli* ATCC 25922 and *CP007799* of *E. coli* Nissle 1917, except for two missing G nucleotides after positions 3245777 and 3245794. We resequenced this fragment and confirmed that the missing nucleotides represent sequencing or genome assembly errors. We named the two correct genes *c3400.1* and *c3400.2*. These genes encode proteins that show 100% sequence identity to *E. coli* TssF and TssG with NCBI references WP_000342476.1 and WP_000373314.1, respectively.

Two PCR product-containing *E. coli* CFT073 genes, *c3402* (TssE) and *c3400.1* (TssF), *c3400.2* (TssG) were fused by overlapping PCR using the primers given in Supplementary Table 4 to create a single construct with *c3402* upstream of *c3400.1* and *c3400.2*. The resulting PCR product was cloned into a modified pBAD33 vector that resembled the pTSL vector described above. TssE carried a TEV protease-cleavable His-SlyD tag at its N terminus. Gene *c3387* (TssK) was cloned into the vector pBAD24 without any tags. Both plasmids were co-transformed into *E. coli* TOP10 cells using appropriate antibiotic selection. These cells were then used for protein expression.

Protocols used for expression and purification of the TssE–TssF–TssG–TssK complex were similar to those used for gp25 with small modifications. The 2 × TY medium was supplemented with ampicillin at 200 mg ml^{-1} and kanamycin at 50 mg ml^{-1}. The plasmid-carrying cells were grown to an A_{600} of 0.8 and induced with arabinose to a final concentration of 0.1% (w/v). The expression was performed at 18 °C with rigorous shaking overnight. The cells were pelleted at 8000*g*, resuspended in lysis buffer (50 mM Tris–HCl pH 8.0, 300 mM NaCl), and disrupted by sonication at 4–16 °C. The insoluble fraction was removed by centrifugation at 27,000*g* for 20 min. The clarified cell lysate was loaded onto a 5-ml GE HisTrap FF Ni-charged column. The protein was eluted with imidazole-containing buffer (50 mM Tris–HCl pH 8.0, 300 mM NaCl, 200 mM imidazole) using a step gradient. Fractions containing the protein were pooled and the TEV protease was added to give a protease/protein ratio of 1/100 (w/w). This mixture was dialysed into 10 mM Tris–HCl pH 8.0 by overnight incubation at 20 °C resulting in cleavage of the His–SlyD expression tag. The digested protein was filtered and purified by ion-exchange chromatography using a MonoQ 10/100 GL column and 0 to 1 M NaCl gradient in 20 mM Tris–HCl pH 8.0 buffer. Fractions containing the TssE–TssF–TssG–TssK complex were pooled and loaded onto a Superdex 200 HiLoad 16/60 size-exclusion column equilibrated with 10 mM Tris-HCl pH 8.0 and 150 mM NaCl.

The TssE–TssF–TssG–TssK complex had an apparent molecular mass of 260 kDa (90% confidence interval is 203–331 kDa) in size-exclusion chromatography on a Superdex 200 10/30 GL column in 10 mM Tris–HCl, pH 8.0, and 150 mM NaCl (Extended Data Fig. 7e, f).

© 2016 Macmillan Publishers Limited. All rights reserved

Construction of T4-7am mutant phage. T4-7am mutant phage was constructed by replacing Ser45 of gene *7* with the amber stop codon by means of homologous recombination with a plasmid carrying this mutation. First, an overlap PCR with primers 7S45AmF, 7S45AmRM, 7S45AmFM and 7S45AmR (Supplementary Table 4) and T4 genomic DNA as a template was performed as described elsewhere[80]. The resulting PCR product was then ligated into the pGem-T vector (Promega) following the manufacturer's recommendations, giving rise to a pG7-45Am plasmid that carried the desired mutation.

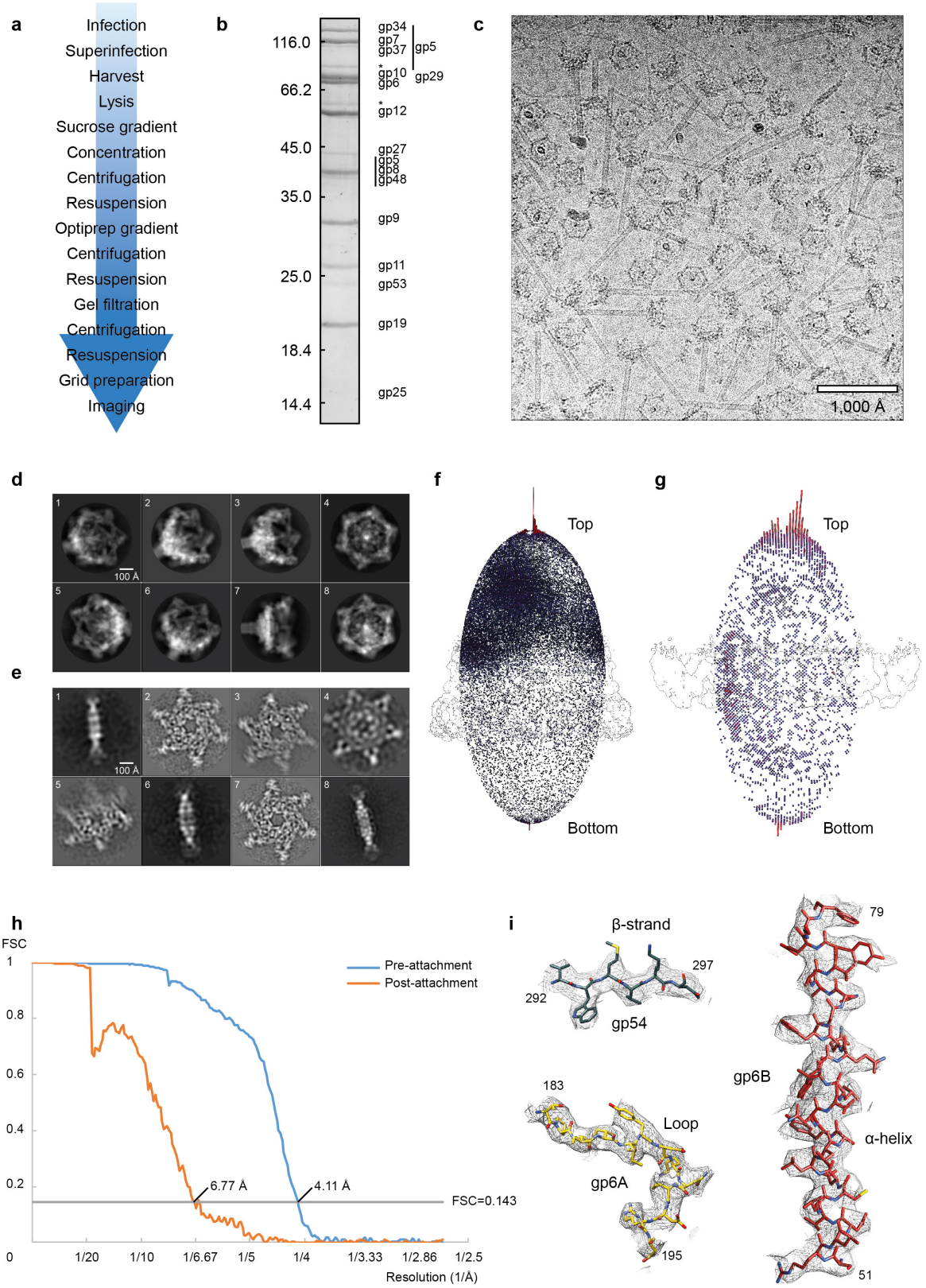
For homologous recombination, pG7-45Am was electroporated into the amber suppressor *supD E. coli* strain CR63 (ref. 81) (serU60(AS)), and the cells were infected with phage T4 with a M.O.I. of 5. After the completion of lysis, the lysate was diluted to obtain an appropriate plaque count when plated onto a CR63 lawn. Several hundred plaques were tested for their ability to grow on CR63 and Sup^0 B^E lawns and a T4-7am mutant phage was found. The presence of the mutation was confirmed by Sanger sequencing.

Complementation assay. The complementation was performed by infecting B^E cells, which expressed the wild-type gp7 protein and its mutants, with the T4-7am mutant phage. First, gene *7* was PCR amplified using T4 genomic DNA as a template and primers T4-7F1 and T4-7R1 (Supplementary Table 4). The PCR fragment was then cloned into the pBAD24 vector to give rise to a plasmid in which gene *7* was under the control of an arabinose promoter. The plasmids for expression of gp7 mutants were constructed by overlap extension with pBAD forward and reverse primers (Invitrogen) as flanking oligonucleotides and the primers given in Supplementary Table 4 as mutagenic oligonucleotides[80]. The wild-type gene *7* plasmid served as a template for all these PCR reactions. The mutations were confirmed by partial sequencing, and the expression levels were analysed by SDS–PAGE[80].

The plasmids expressing all variants of gp7 and an empty pBAD24 vector (for negative control) were transformed into the B^E strain by electroporation. Night cultures were diluted 100 times into an LB medium supplemented with ampicillin (Sigma) at 100 mg ml^{-1} and then incubated at 37 °C with shaking at 220 r.p.m. When the cultures reached an A_{600} of 0.150 they were induced by the addition of arabinose (Sigma) to a concentration of 0.02% (w/v). When the cells reached an A_{600} of 0.600, T4-7am was added with a M.O.I. of 5. Superinfection was performed 10 min later at the same M.O.I. After 4 h, a few drops of chloroform were added to complete cell lysis. The phages were harvested and purified by centrifugation in a stepped gradient of caesium chloride[82] and then dialysed into SM buffer (50 mM TrisCl pH 7.5, 100 mM NaCl, 8 mM $MgSO_4$). The concentration of the samples was normalized according to their absorbance at 260 nm and then examined using SDS–PAGE (Extended Data Fig. 9b). The infectivity was tested by plaque assay on *E. coli* CR63 (Extended Data Fig. 9c) and *E. coli* B^E (for negative control).

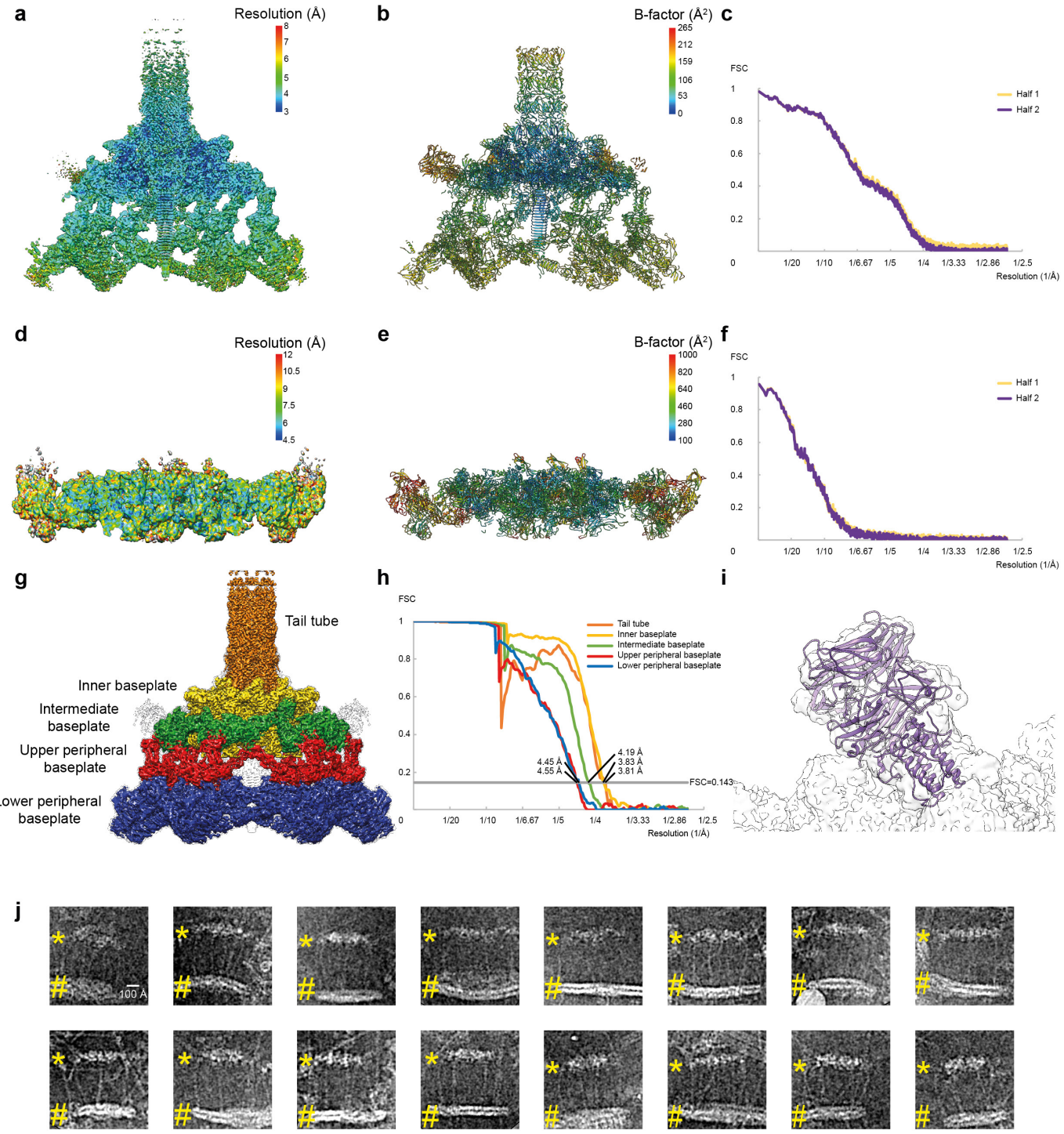
51. Doermann, A. H. Lysis and lysis inhibition with *Escherichia coli* bacteriophage. *J. Bacteriol.* **55,** 257–276 (1948).
52. Li, X. *et al.* Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nature Methods* **10,** 584–590 (2013).
53. Scherer, S. *et al.* 2dx_automator: implementation of a semiautomatic high-throughput high-resolution cryo-electron crystallography pipeline. *J. Struct. Biol.* **186,** 302–307 (2014).
54. Rohou, A. & Grigorieff, N. CTFFIND4: fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192,** 216–221 (2015).
55. Tang, G. *et al.* EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157,** 38–46 (2007).
56. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180,** 519–530 (2012).
57. Rosenthal, P. B. & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.* **333,** 721–745 (2003).
58. Pintilie, G. D., Zhang, J., Goddard, T. D., Chiu, W. & Gossard, D. C. Quantitative analysis of cryo-EM density map segmentation by watershed and scale-space filtering, and fitting of structures by alignment to regions. *J. Struct. Biol.* **170,** 427–438 (2010).
59. Leiman, P. G. *et al.* Structure of bacteriophage T4 gene product 11, the interface between the baseplate and short tail fibers. *J. Mol. Biol.* **301,** 975–985 (2000).
60. Leiman, P. G., Shneider, M. M., Mesyanzhinov, V. V. & Rossmann, M. G. Evolution of bacteriophage tails: Structure of T4 gene product 10. *J. Mol. Biol.* **358,** 912–921 (2006).
61. Leiman, P. G. *et al.* Structure and location of gene product 8 in the bacteriophage T4 baseplate. *J. Mol. Biol.* **328,** 821–833 (2003).
62. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
63. Söding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33,** W244–W248 (2005).
64. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols* **10,** 845–858 (2015).
65. Eswar, N. *et al.* Comparative protein structure modeling using MODELLER. *Curr. Protoc. Protein Sci.* **50,** 2.9:2.9.1–2.9.31 (2007).
66. Perrakis, A., Morris, R. & Lamzin, V. S. Automated protein model building combined with iterative structure refinement. *Nature Struct. Biol.* **6,** 458–463 (1999).
67. Afonine, P. V., Headd, J. J., Terwilliger, T. C. & Adams, P. D. New tool: phenix. real_space_refine. *Comput. Crystallogr. Newslett.* **4,** 43–44 (2013).
68. Brown, A. *et al.* Tools for macromolecular model building and refinement into electron cryo-microscopy reconstructions. *Acta Crystallogr. D* **71,** 136–153 (2015).
69. Kabsch, W. Solution for best rotation to relate 2 sets of vectors. *Acta Crystallogr. A* **32,** 922–923 (1976).
70. Collaborative Computational Project Number 4. The *CCP*4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50,** 760–763 (1994).
71. Tao, Y., Strelkov, S. V., Mesyanzhinov, V. V. & Rossmann, M. G. Structure of bacteriophage T4 fibritin: a segmented coiled coil and the role of the C-terminal domain. *Structure* **5,** 789–798 (1997).
72. Abuladze, N. K., Gingery, M., Tsai, J. & Eiserling, F. A. Tail length determination in bacteriophage T4. *Virology* **199,** 301–310 (1994).
73. Han, K. Y. *et al.* Solubilization of aggregation-prone heterologous proteins by covalent fusion of stress-responsive *Escherichia coli* protein, SlyD. *Protein Eng. Des. Sel.* **20,** 543–549 (2007).
74. Leslie, A. G. W. & Powell, H. R. in *Evolving Methods for Macromolecular Crystallography* (eds Read, R. J. & Sussman, J. L.) vol. 245, 41–51 (Springer, 2007).
75. Evans, P. Scaling and assessment of data quality. *Acta Crystallogr. D* **62,** 72–82 (2006).
76. Evans, P. R. An introduction to data reduction: space-group determination, scaling and intensity statistics. *Acta Crystallogr. D* **67,** 282–292 (2011).
77. Sheldrick, G. M. A short history of SHELX. *Acta Crystallogr. A* **64,** 112–122 (2008).
78. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
79. Brunet, Y. R., Zoued, A., Boyer, F., Douzi, B. & Cascales, E. The type VI secretion TssEFGK-VgrG phage-like baseplate is recruited to the TssJLM membrane complex via multiple contacts and serves as assembly platform for tail tube/sheath polymerization. *PLoS Genet.* **11,** e1005545 (2015).
80. Sambrook, J. & Russell, D. W. *Molecular Cloning: a Laboratory Manual* (Cold Spring Harbor Laboratory Press, 2001).
81. Appleyard, R. K., McGregor, J. F. & Baird, K. M. Mutation to extended host range and the occurrence of phenotypic mixing in the temperate coliphage lambda. *Virology* **2,** 565–574 (1956).
82. Clokie, M. R. J. & Kropinski, A. M. *Bacteriophages: Methods and Protocols,* (Humana Press, 2009).
83. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nature Methods* **11,** 63–65 (2014).
84. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14,** 1188–1190 (2004).
85. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372,** 774–797 (2007).

© 2016 Macmillan Publishers Limited. All rights reserved



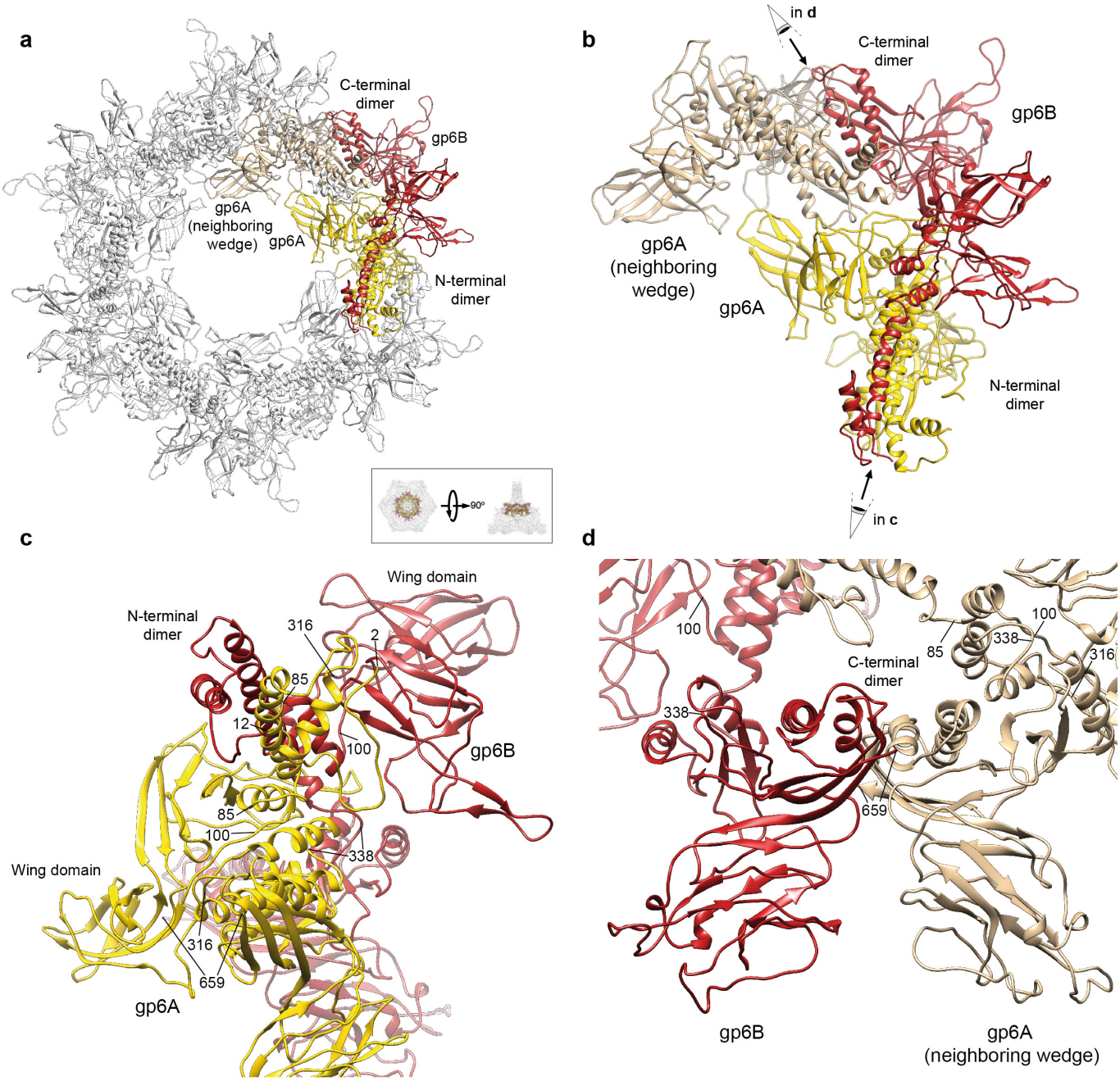
Extended Data Figure 1 | Sample purification, cryo-EM imaging and reconstruction. a, Purification scheme of the T4 baseplate–tail tube complexes. **b**, SDS–PAGE of the sample used in cryo-EM imaging (the full gel is shown in Supplementary Fig. 1). **c**, Raw cryo-EM image of T4 baseplates. **d**, **e**, Representative reference-free 2D class averages of T4 baseplates in pre- and post-attachment conformations, respectively. The number of particles in each class in **d** is as follows: 1: 534; 2: 813; 3: 689; 4: 655; 5: 292; 6: 858; 7: 2,223; 8: 977. The number of particles in each class in **e** is as follows: 1: 315; 2: 68; 3: 494; 4: 60; 5: 263; 6: 387; 7: 62; 8: 566. **f**, **g**, Distribution of refined angles of the baseplate in both conformations. **h**, Fourier shell correlation (FSC) between independently refined maps calculated using half of the data (gold-standard refinement) after post-processing for both conformations. **i**, Fragments of the pre-attachment baseplate cryo-EM reconstruction map with fitted atomic model.

© 2016 Macmillan Publishers Limited. All rights reserved



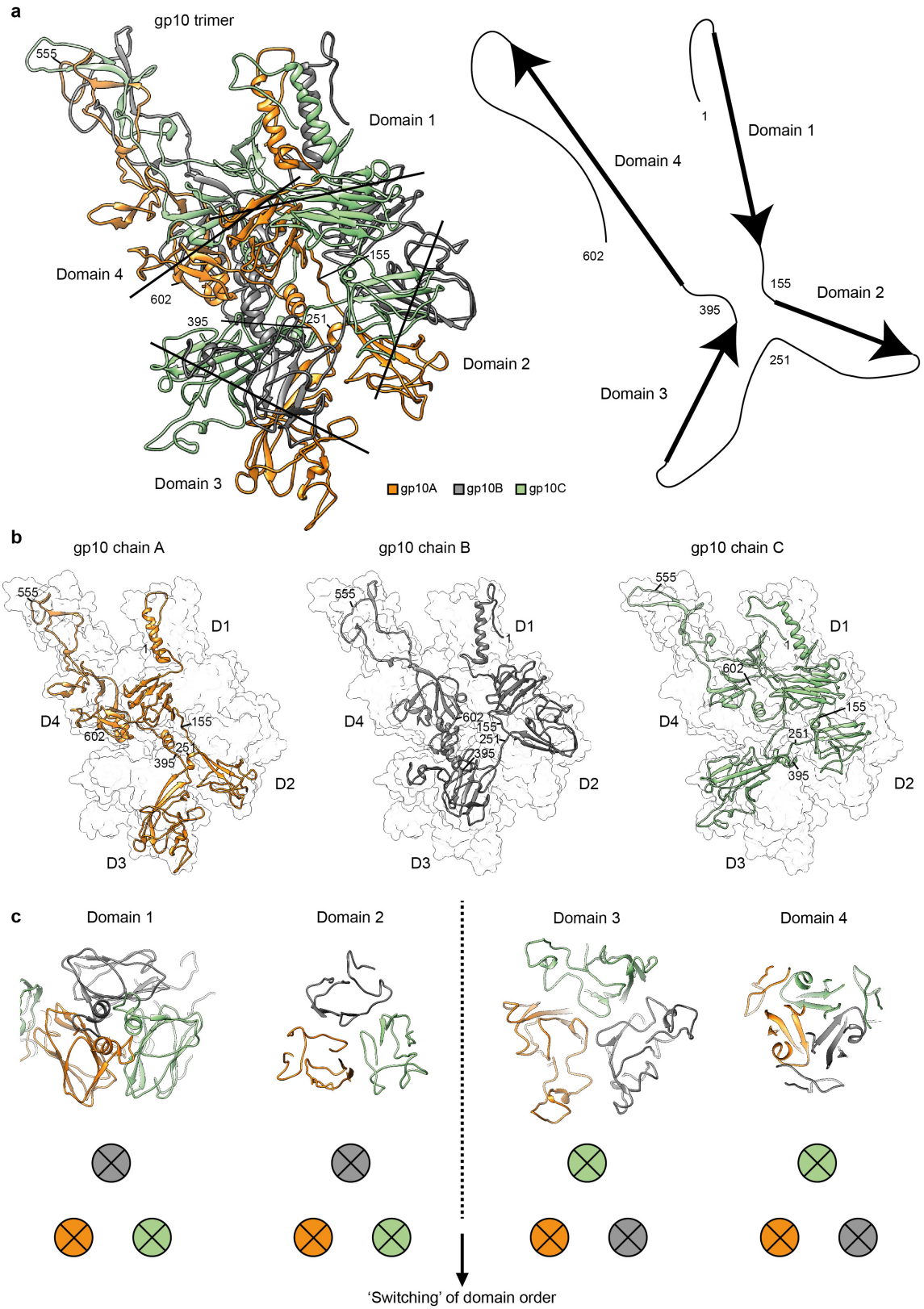
Extended Data Figure 2 | Details of local resolution estimation and baseplates attached to membranes. a, **d**, Resolution of pre- and post-attachment reconstructions analysed with ResMap[83]. **b**, **e**, Atomic models of pre- and post-attachment baseplates coloured by B-factors. **c**, Plots of the model-map FSC of the scrambled pre-attachment baseplate structure refined against half data map 1 from the gold-standard refinement versus half data map 1 (the map it was refined against) and half data map 2 (against which it was not refined). The absence of a large gap between both curves indicates that no excessive overfitting took place. **f**, As in **c**, but for the post-attachment baseplate. **g**, Localized cryo-EM refinement maps of the pre-attachment baseplate. **h**, FSC graphs of the localized refinements in **g**. **i**, Fit of gp9 into the pre-attachment baseplate map (the contour level is lower than that in Fig. 1a). **j**, Gaussian low-pass filtered (1/25 Å) raw images of individual baseplates attached to cell membranes selected from $n = 243$ similar images. Asterisks, baseplates; hash symbols, membranes. The extended STFs connecting the two can be seen.

© 2016 Macmillan Publishers Limited. All rights reserved



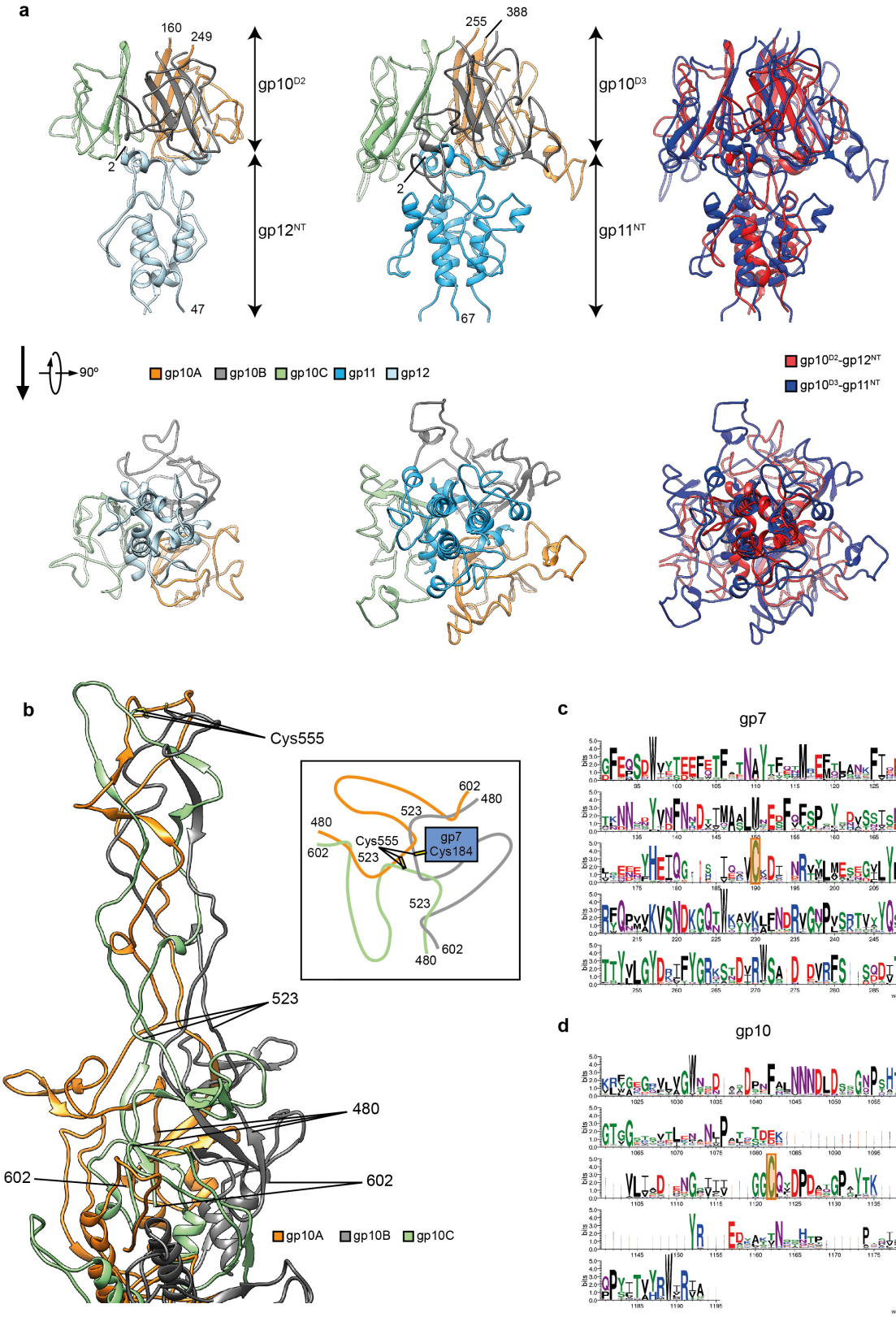
Extended Data Figure 3 | Structure of the gp6 ring. a, Two structurally non-identical copies comprising the asymmetric unit of the ring are coloured yellow (gp6A) and red (gp6B). A neighbouring gp6A copy is coloured light brown. The rest of the ring is shown in light grey. The inset shows the position of the ring in the baseplate map. **b**, A close-up view of the gp6 subunits coloured in **a**. **c**, **d**, Close-up views of the N- and C-terminal gp6 dimers, respectively. In both views, the six-fold axis of the baseplate is roughly vertical.

© 2016 Macmillan Publishers Limited. All rights reserved



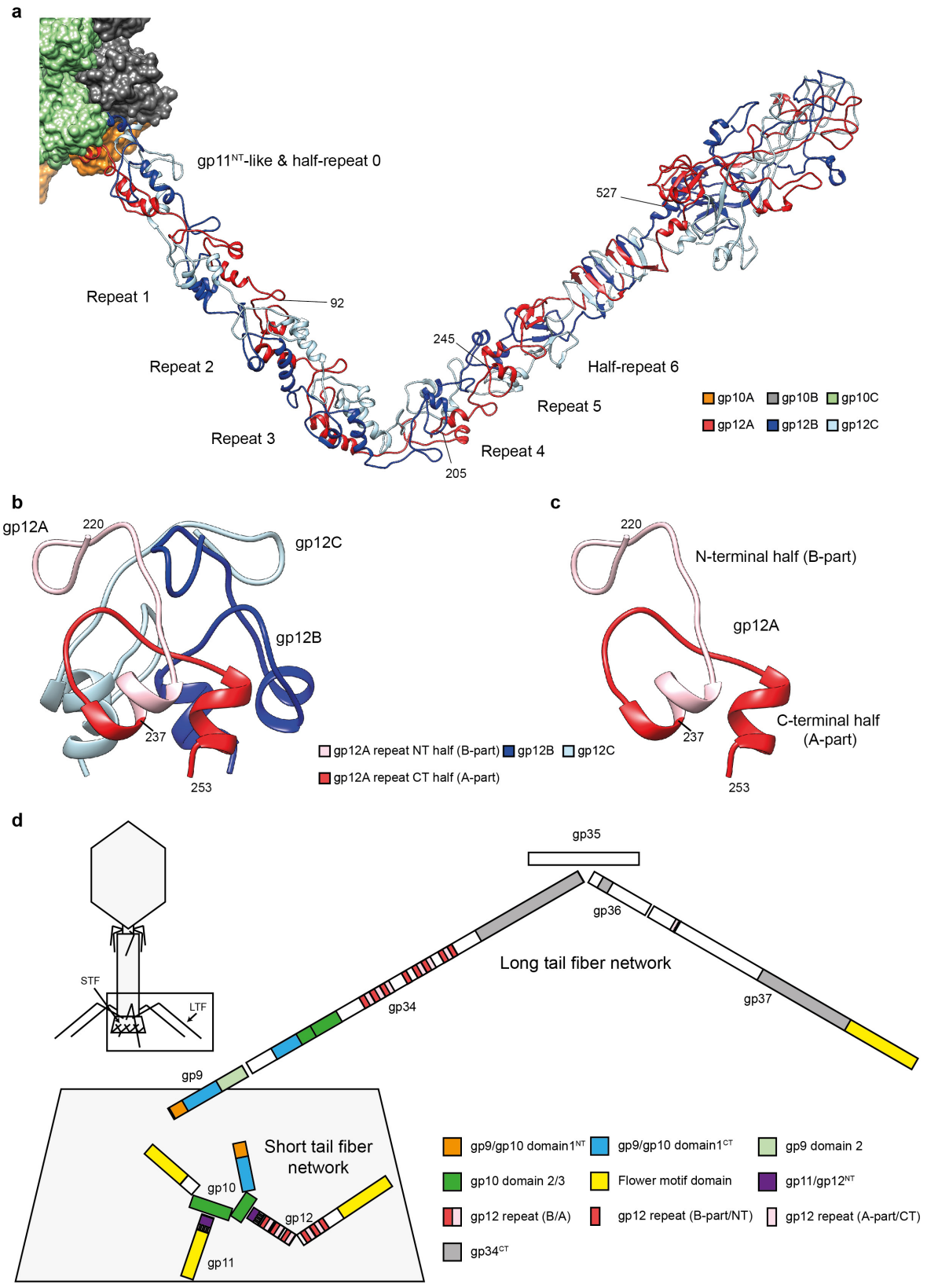
Extended Data Figure 4 | Structure of gp10. a, Ribbon diagram of gp10 trimer and a schematic showing the general direction of the polypeptide chain in its four domains. The arrows also indicate the viewing direction in **c**. **b**, Traces of the three chemically identical polypeptide chains making up the complete gp10 molecule, which is shown as a semitransparent surface. **c**, Cross-sections through the four domains of gp10 in positions indicated with black lines in **a**. Note the switch of chain order around the three-fold axis between domains 2 and 3.

© 2016 Macmillan Publishers Limited. All rights reserved



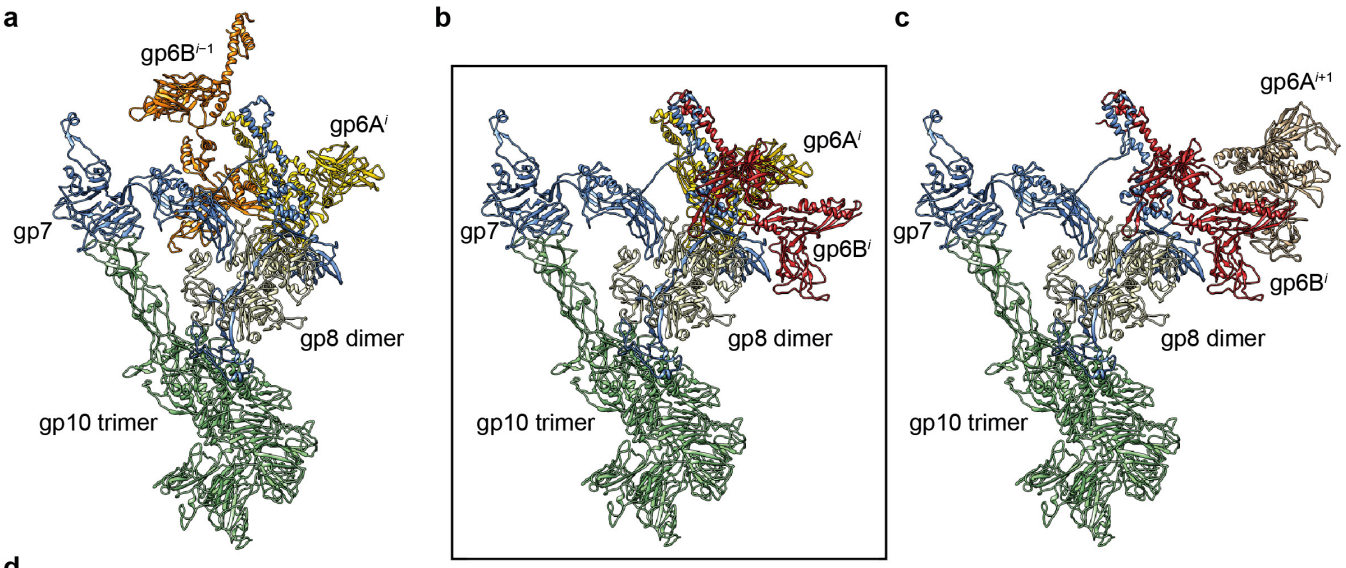
Extended Data Figure 5 | Interaction of gp10 with gp11 and gp12. **a**, Structure and binding of gp10 domains 2 and 3 to the N-terminal domains of gp12 and gp11, respectively. The rightmost panel shows a superposition of the two complexes. **b**, The structure of gp10 domain 4. Cys555 and several other residues at strategic locations are labelled. Inset: a diagram explaining the complex knotted topology of gp10 domain 4. The gp10–gp10 and gp10–gp7 inter-chain disulfide bridges are indicated. **c**, **d**, Conservation of gp7 and gp10 amino acid sequences, respectively, is represented in the WebLogo format with letter heights proportional to the degree of conservation[84]. The conserved cysteines are highlighted with orange boxes. The numbers below the letters are positions in a multiple sequence alignment.

© 2016 Macmillan Publishers Limited. All rights reserved



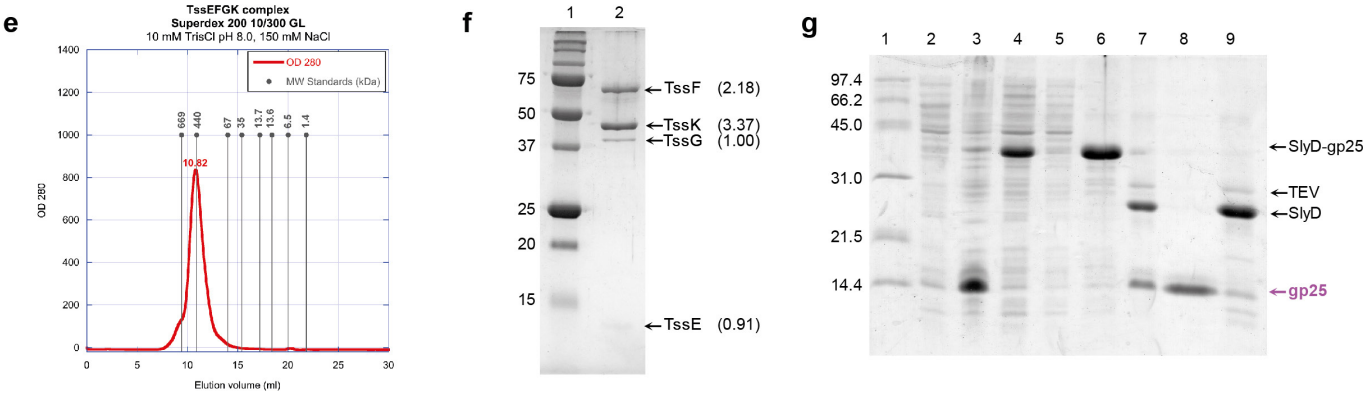
Extended Data Figure 6 | Structure of gp12 and domain organization of T4 fibres. a, Ribbon diagram of the gp12 trimer, anchored to gp10 domain 2 (shown in surface representation). The N-terminal part of the fibre (residues 2–245) was built *de novo*. **b**, Structure of the gp12 repeat. **c**, Fold of the polypeptide chain making up the repeat. **d**, Evolutionary relationships between different T4 proteins comprising the baseplate's periphery and fibres. The size of each bar is proportional to the amino acid sequence length. The gp12 repeat shown in **b** and **c** constitutes a major part of the proximal LTF protein gp34.

© 2016 Macmillan Publishers Limited. All rights reserved



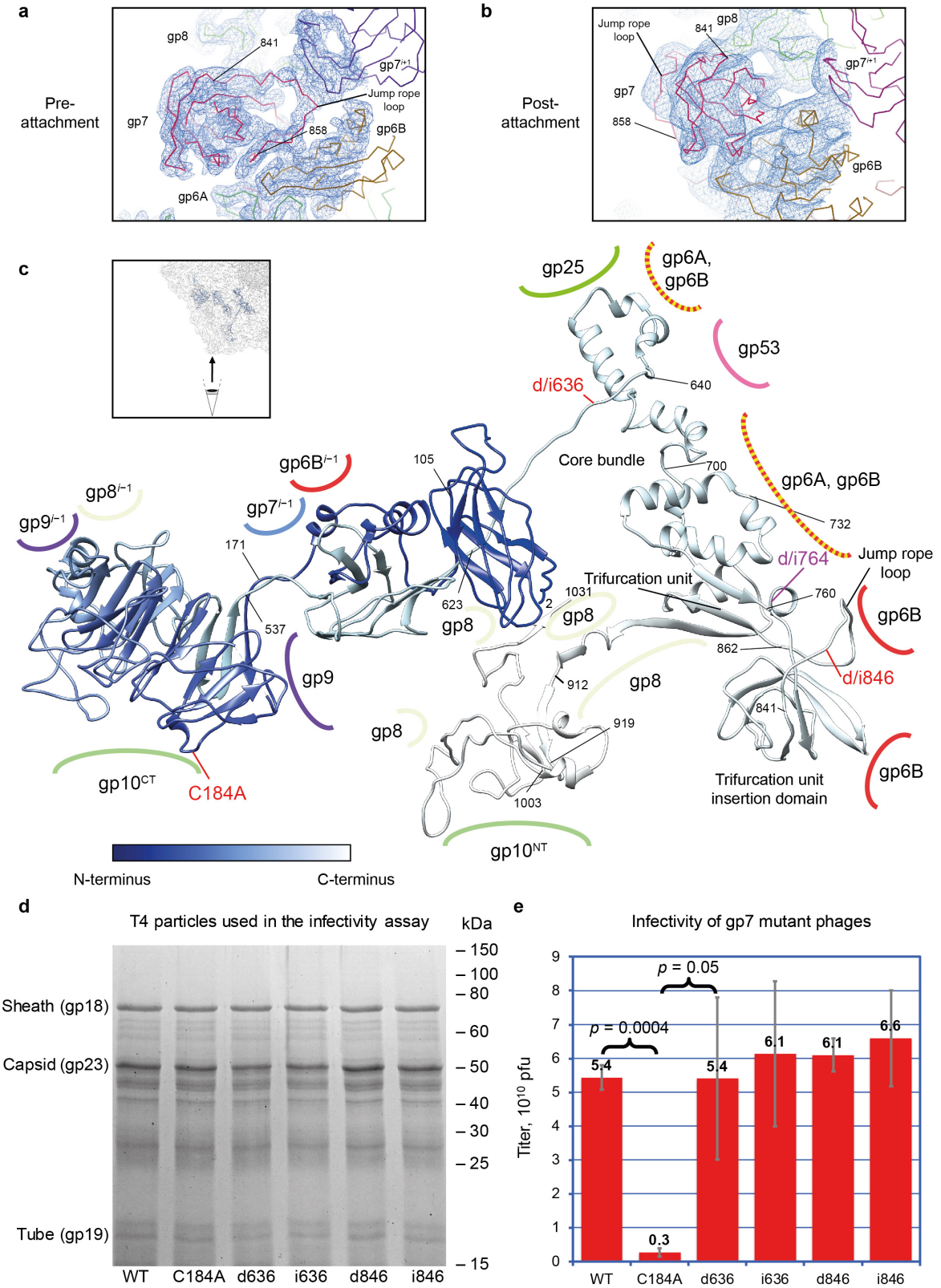
d

	Potential wedge intermediate A	Potential wedge intermediate B	Potential wedge intermediate C
BSA (Å^2)	73,018.2	81,929.6	72,413.9
ΔG^0_{diss} (kcal/mol)	-2.3	36.0	-6.3



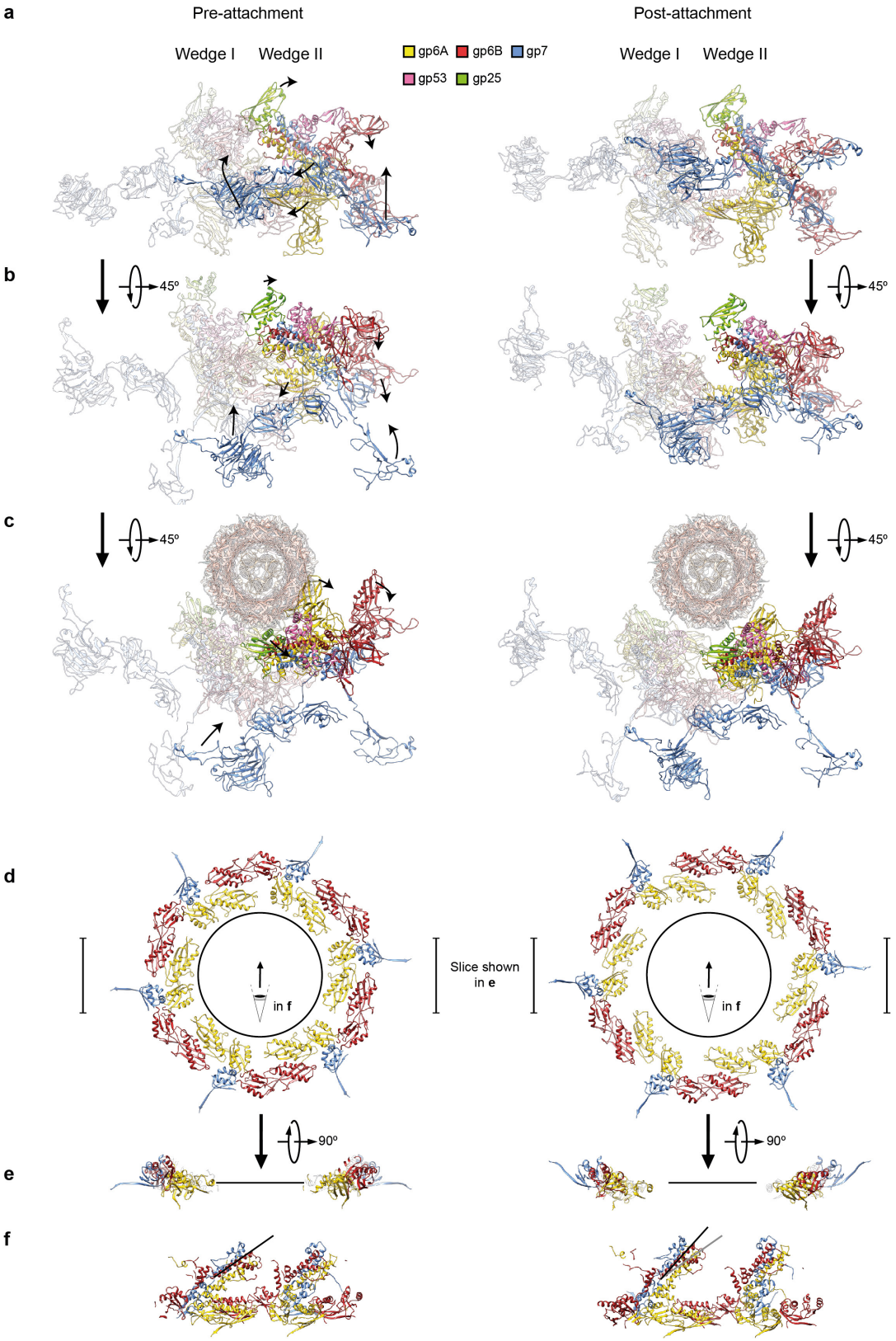
Extended Data Figure 7 | T4 and T6SS baseplate assembly. Determination of the structure of the wedge precursor on biophysical grounds (see Supplementary Information). **a**–**c**, Three possible precursors with a composition of $(gp6)_2$–gp7–$(gp8)_2$–$(gp10)_3$ and association as found in the pre-attachment baseplate. **d**, Calculations with PISA[85] favour the assembly shown in **b**. **e**, Size-exclusion chromatography of $(TssE)_1$–$(TssF)_2$–$(TssG)_1$–$(TssK)_3$ complex. **f**, 12% SDS–PAGE of $(TssE)_1$–$(TssF)_2$–$(TssG)_1$–$(TssK)_3$ complex. Numbers on the left are MW standards in kDa. Numbers in parentheses to the right of protein names are relative band intensities as quantified by Image Studio Lite (LI-COR) (the full gel is shown in Supplementary Fig. 1). **g**, Expression and purification of T4 gp25 in soluble form. Fraction 8 contains pure gp25 that was used for crystallization and structure determination. See Methods.

© 2016 Macmillan Publishers Limited. All rights reserved



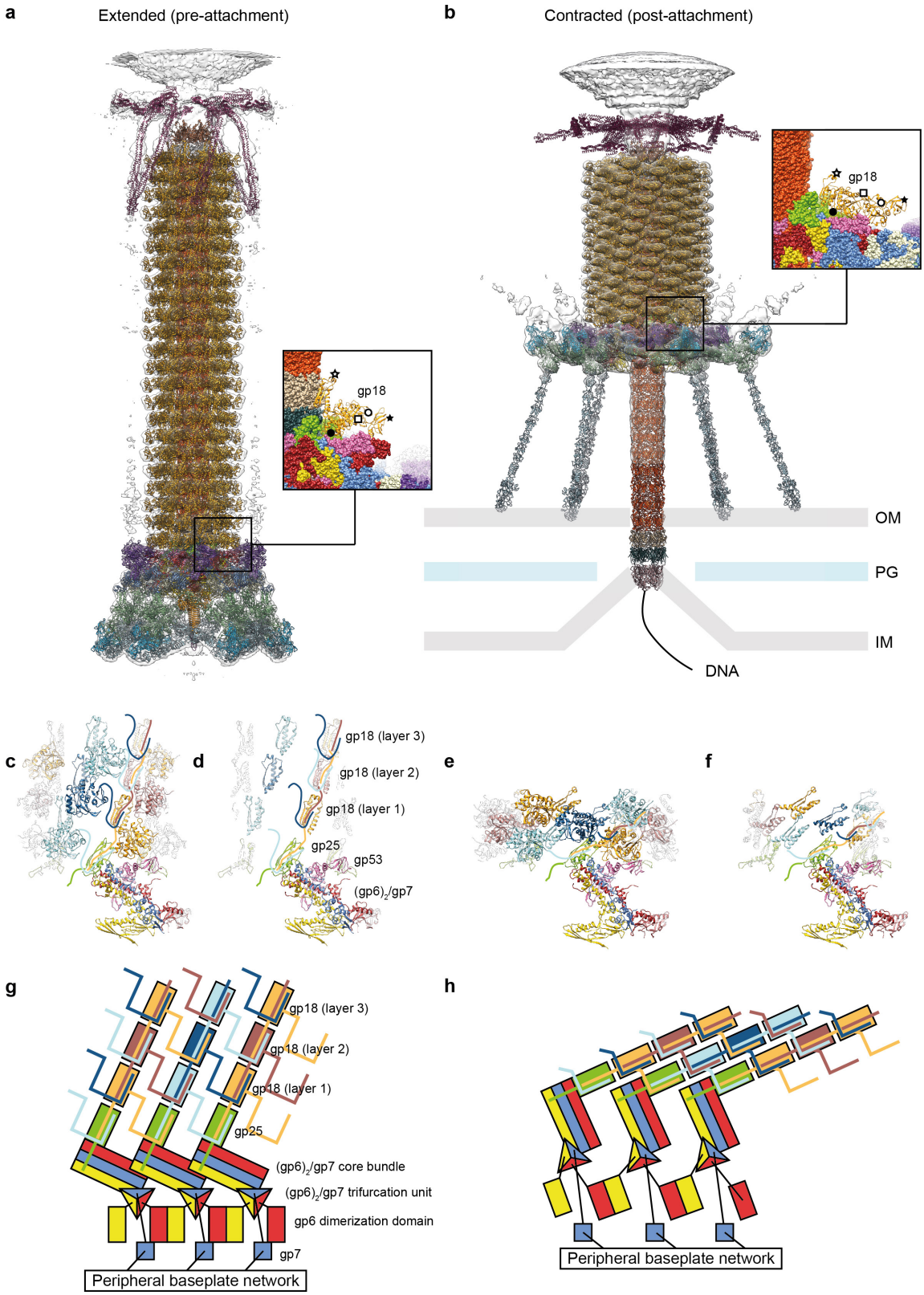
Extended Data Figure 8 | Contacts of gp7 with other baseplate proteins and gp7 mutagenesis. a, **b**, Location of the gp7 jump rope loop in the two conformations of the baseplate. **c**, The structure of gp7 is coloured by amino acid number from blue N terminus to white C terminus. The inset shows the position of gp7 (coloured blue) within the baseplate map. Interactions of gp7 with other baseplate proteins are shown schematically using the colour code of Fig. 1. The superscript $i-1$ denotes interactions with a symmetry-related copy of a given protein. Red letters indicate sites of mutagenesis (for example, d/i636: deletion/insertion at position 636) that resulted in viable phage particles (see Supplementary Information). The purple label shows the site that tolerated neither residue deletion nor insertion. **d**, SDS–PAGE of purified T4-7am particles carrying the wild-type (WT) and five mutant gp7 proteins after their concentrations were brought to a common scale according to their absorbance at 260 nm (the full gel is shown in Supplementary Fig. 1). **e**, Infectivity of the mutant particles shown in **d**. The error bar indicates the 95% confidence interval obtained in three independent experiments ($n = 3$). Only one mutation was statistically significantly different from the rest and *P*-values (two-tailed Student's t-test) comparing it to the wild-type phage and to the mutant with the largest experimental error are given.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 9 | Transformation of the conserved part of the baseplate. a–c, Three different views of the inner and intermediate parts of two adjacent wedges (gp25, gp53, gp6 and gp7). Complete proteins are shown, including weakly conserved domains. The left and right columns represent the pre- and post-attachment baseplates, respectively. One of the two wedges (wedge I) is semi-transparent. In **c**, the central spike–tail tube complex is also displayed (semi-transparent). **d**, **e**, Transformation of the ring of the $(gp6)_2$–gp7 trifurcation and gp6 dimerization domains from the pre- to the post-attachment state (as in Fig. 2d). The bars in **d** indicate the size of the fragment shown in **e** after rotation. **f**, Reorientation of the $(gp6)_2$–gp7 core bundles.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 10 | Model for baseplate-induced sheath contraction. a, **b**, Pseudoatomic model of the complete T4 tail in the extended (pre-attachment) and contracted (post-attachment) conformations. The insets show a close-up view (labelled with a black box) of the position of the gp18 subunit on the baseplate in the extended and contracted conformations of the sheath. The white geometrical shapes label the same regions of the sheath subunit in both conformations. **c**, Interaction of the two conserved domains of the gp18 sheath protein with the conserved components of the T4 baseplate wedge (as in Fig. 3). Coloured lines indicate the putative topology of the N- and C-terminal gp18 extensions, as well as the gp25 C-terminal strand. **d**, The same view as in **c**, but the external domain is now not shown for clarity to demonstrate the interaction of gp25-like sheath domains with each other and with gp25. **e**, **f**, The same as **c** and **d** but in the contracted state. **g**, **h**, Two diagrams demonstrating the motion of baseplate components that results in sheath contraction.

© 2016 Macmillan Publishers Limited. All rights reserved

ARTICLE

doi:10.1038/nature17990

Transcription initiation complex structures elucidate DNA opening

C. Plaschka[1]*, M. Hantsche[1]*, C. Dienemann[1], C. Burzinski[1], J. Plitzko[2] & P. Cramer[1]

Transcription of eukaryotic protein-coding genes begins with assembly of the RNA polymerase (Pol) II initiation complex and promoter DNA opening. Here we report cryo-electron microscopy (cryo-EM) structures of yeast initiation complexes containing closed and open DNA at resolutions of 8.8 Å and 3.6 Å, respectively. DNA is positioned and retained over the Pol II cleft by a network of interactions between the TATA-box-binding protein TBP and transcription factors TFIIA, TFIIB, TFIIE, and TFIIF. DNA opening occurs around the tip of the Pol II clamp and the TFIIE 'extended winged helix' domain, and can occur in the absence of TFIIH. Loading of the DNA template strand into the active centre may be facilitated by movements of obstructing protein elements triggered by allosteric binding of the TFIIE 'E-ribbon' domain. The results suggest a unified model for transcription initiation with a key event, the trapping of open promoter DNA by extended protein–protein and protein–DNA contacts.

For transcription initiation, Pol II assembles with the basal transcription factors (TF) IIB, TFIID (or its subunit TBP), TFIIE, TFIIF, and TFIIH[1–4] on double-stranded promoter DNA to form the closed complex (CC). Upon DNA opening, the template strand slips into the Pol II active centre, and a DNA 'bubble' forms ~20–30 base pairs (bp) downstream of the TATA box[5], leading to the open complex (OC). Efficient DNA opening requires TFIIE and TFIIH[6–8], but these factors are not required for low levels of transcription[9,10] and TFIIE alone can open certain promoters[11,12]. Subsequent RNA synthesis results in the initially transcribing complex (ITC), which is then converted into an elongation complex for processive RNA synthesis.

The three-dimensional architecture of Pol II initiation complex intermediates was studied in yeast[13–19] and human[20] systems. These studies revealed that the promoter assembly containing TBP and TFIIB resides over the Pol II wall and positions DNA above the polymerase cleft and along the clamp. TFIIE and TFIIF bind to opposite sides of Pol II and flank promoter DNA. Owing to the limited resolution of available structural studies, many questions regarding the molecular basis of initiation remain, including the mechanisms of DNA opening and template-strand loading into the active centre. A differing model of the yeast initiation complex[17] was revised recently[21], and now agrees, on a topological level, with other studies and results reported here.

Here we report cryo-EM structures of CC and OC assemblies from the yeast *Saccharomyces cerevisiae* at resolutions of 8.8 Å and 3.6 Å, respectively. The structures contain all of the basal transcription factors except TFIIH, which is currently not available at a quality required for high-resolution analysis. We show that DNA opening can occur in the absence of TFIIH, and provide mechanistic insights into DNA opening and template-strand loading. Our results also unveil the exact location and intricate structure of basal factors and their induced folding interactions with each other, with Pol II, and with promoter DNA. The data also demonstrate the high structural conservation between yeast and human initiation systems.

Transcription initiation complex at 3.6 Å

We extended our previous cryo-EM analysis of a Pol II core initiation complex containing TBP, TFIIB, and TFIIF[19] by adding TFIIA and TFIIE (Methods). Formation of a stable and stoichiometric complex required the presence of core Mediator, which showed low occupancy and high flexibility under cryo-EM conditions, as observed previously[19], and was excluded from further analysis (Methods, Extended Data Fig. 1a–c). We acquired 257,259 cryo-EM single particle images using a K2 direct electron detector (Methods, Extended Data Fig. 1b, c). Unsupervised particle sorting led to an OC structure at an overall resolution of 3.6 Å, revealing the Pol II core at up to 3.1 Å (OC1), compared to 7.8 Å for our core ITC structure[19] (Fig. 1a, b, Extended Data Figs 1d–g). Further particle sorting revealed improved density for TFIIB and TFIIF at around 4 Å resolution (OC2 and OC4), and for TFIIE at 4.4 Å resolution (OC3, Methods, Extended Data Figs 1i, k, l). The final structure mainly consists of atomic models (90%) and contains backbone models for parts of the basal factors (Extended Data Fig. 1f–l).

The structure reveals upstream DNA above the Pol II wall and downstream DNA in the active centre cleft[19]. We observed only fragmented density for single strands within the DNA bubble (Fig. 1b, c, Extended Data Fig. 1m), and no density for RNA, which was not retained during cryo-EM sample preparation (not shown). The bubble contains 15 mismatches, begins at the natural distance of ~20 bp downstream of the TATA box[5], and extends further downstream than when it initially forms, resembling the situation during transcription start-site scanning. For the basal factors, regions that are essential for cell viability are generally observed, whereas non-essential, non-conserved regions[16,22–26] are often mobile (Fig. 1a, b, Extended Data Fig. 1 j–l). TFIIA is located near upstream DNA and TBP as expected[23] (Fig. 1b, Extended Data Fig. 1j). For TFIIB, the B-ribbon and B-core domains are well defined, including the newly modelled carboxy-terminal cyclin domain from yeast (Fig. 2), whereas the B-linker shows weak density, and the B-reader[15,19] is mobile (Extended Data Fig. 1i).

DNA positioning and retention

TFIIF adopts an intricate fold within the OC. Its dimerization module and charged helix are located on the Pol II lobe domain as in the ITC[18,19] (Extended Data Fig. 2). The 'arm' in the large TFIIF subunit Tfg1 (human RAP74) adds a β-strand to the Pol II protrusion, and projects into the cleft, where it may stabilize the DNA bubble[18–20]

[1]Max Planck Institute for Biophysical Chemistry, Department of Molecular Biology, Am Fassberg 11, 37077 Göttingen, Germany. [2]Max Planck Institute for Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany.
*These authors contributed equally to this work.

© 2016 Macmillan Publishers Limited. All rights reserved

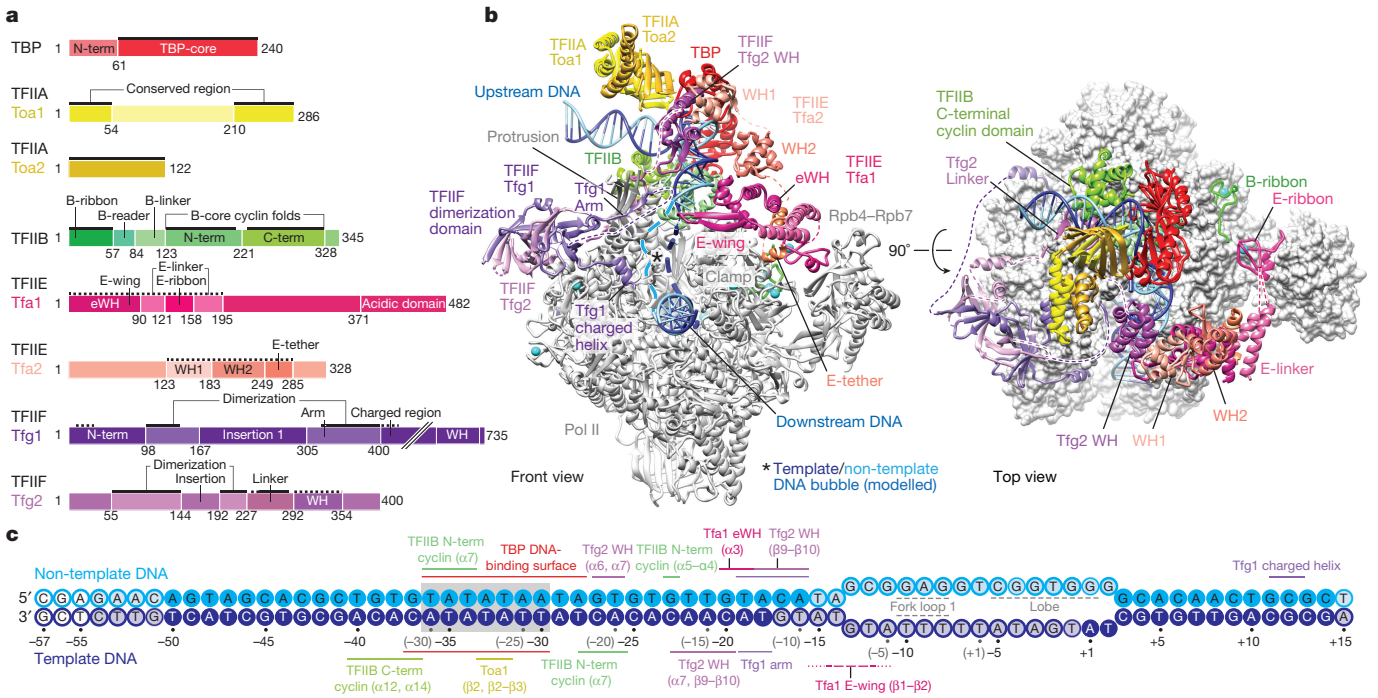


Figure 1 | Open complex structure at 3.6 Å resolution. a, Domain organization of yeast basal transcription factors TBP, TFIIA, TFIIB, TFIIE, and TFIIF. Solid and dashed black bars indicate protein regions that are present in the OC structure as atomic and backbone models, respectively. Colour code used throughout. **b**, Two views[50] of the yeast OC structure. Pol II is in silver. DNA template and non-template strands are in dark blue and cyan, respectively. On the right, Pol II is shown as a surface representation, all other proteins are shown as ribbon models. **c**, Protein–DNA contacts. Promoter DNA nucleotides are depicted with solid, shaded, and empty circles when they were included in the structure, excluded owing to weak density, or excluded owing to a lack of density, respectively. Solid and dashed lines indicate observed and putative protein interactions, respectively. A magenta dashed line indicates the contact between closed DNA and the TFIIE E-wing. The register of promoter DNA is given for analogous yeast (black) and human (grey) positions with respect to the transcription start site (TSS, +1). TATA box is indicated by a grey box.

(Fig. 2). The linker in TFIIF subunit Tfg2 (human RAP30) emanates from the dimerization module, and winds along the base of the protrusion, where it binds a hydrophobic pocket. The Tfg2 linker continues between the protrusion and the TFIIB cyclin domains, and connects to the Tfg2 C-terminal winged-helix (WH) domain on top of the cleft (Fig. 2, Extended Data Fig. 2f). The Tfg2 linker stabilizes TFIIB on the wall of Pol II[27,28]. The yeast-specific amino-terminal region of Tfg1 binds near the Pol II external 1 domain, according to a separate crystallographic analysis (Extended Data Fig. 2d, e, Extended Data Fig. 9c).

TFIIE is located between the clamp and the Rpb4–Rpb7 stalk (Fig. 3a, b, Methods, and Extended Data Fig. 3), consistent with previous topological placement of TFIIE[16,20,29–31] and its archaeal counterpart[32]. The TFIIE structure differs from a recent model obtained at 6 Å resolution[21] (see Extended Data Fig. 1e legend for details, Extended Data Fig. 3, and refs 16, 17). The large TFIIE subunit Tfa1 (human TFIIEα) contains an extended winged helix ('eWH') domain[33] and a zinc ribbon domain[34] ('E-ribbon') that are connected by α-helices (called here 'E-linker') (Fig. 3b). The eWH domain uses its helix α3 to contact both the tip of the Pol II clamp helices and the DNA backbone at positions −13/−14 upstream of the transcription start site (TSS, position +1) (Fig. 1c, 3a). The E-ribbon binds between the clamp, the Rpb7 oligonucleotide-binding (OB) domain, and the B-ribbon (Fig. 3a). The E-linker[20,21] and the mobile C-terminal domain of Tfa1 contact TFIIH[7,35], which may alter TFIIE conformation. The small TFIIE subunit Tfa2 (human TFIIEβ) contains two WH domains ('WH1' (ref. 36), 'WH2'), and two conserved α-helices (called here 'E-tether') that bind the E-linker (Fig. 3b). Consistent with the structure, the E-tether is essential for TFIIE function[16,37]

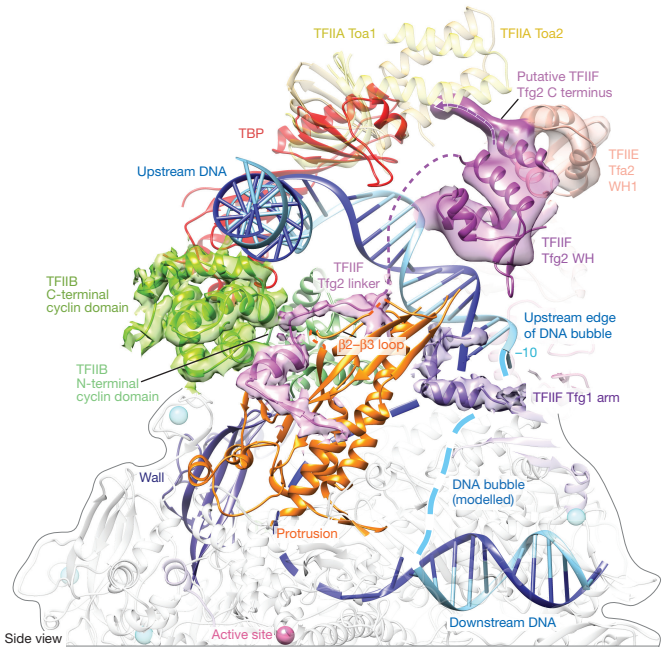


Figure 2 | Basal factors position and retain DNA. Details of the upstream DNA assembly viewed from the side[50]. Highlighted are the locations of the Pol II wall (navy blue), protrusion (orange), TBP (red), TFIIB (green), TFIIF arm (purple), Tfg2 linker (dark magenta) and winged helix (light purple), TFIIE Tfa2 WH1 (light salmon), DNA (template, blue; non-template, cyan), and the active site (magenta sphere). TFIIA is transparent (light yellow). Cryo-EM densities for newly modelled regions of TFIIB, TFIIF, and TFIIE are superimposed on their structural models. Interactions with the Pol II protrusion and upstream edge of the DNA bubble are indicated. TBP contacts a density assigned to the Tfg2 C-terminal region, consistent with interaction of their human counterparts[29] (Extended Data Fig. 2f).

© 2016 Macmillan Publishers Limited. All rights reserved

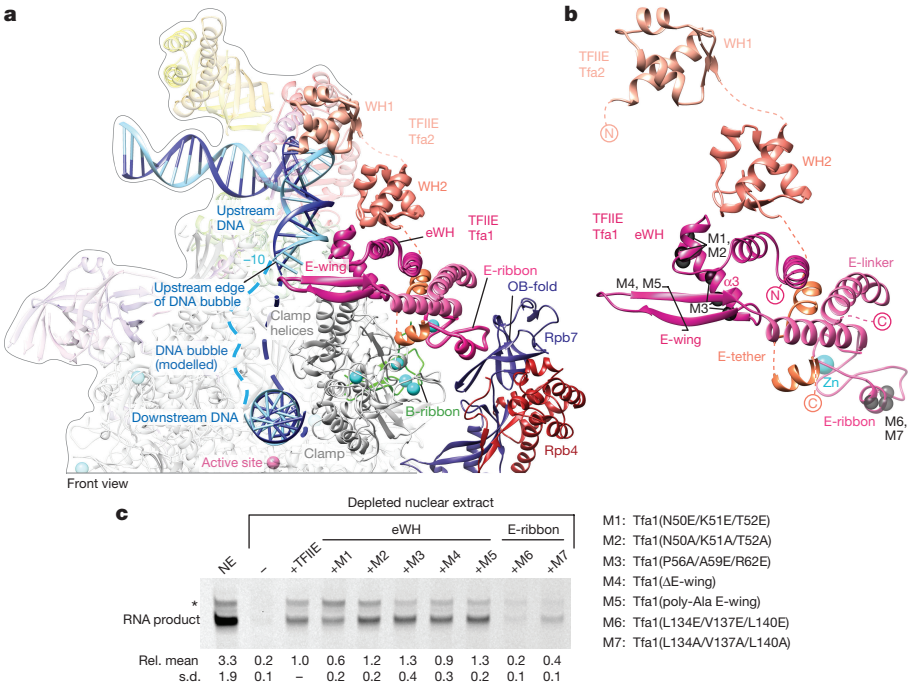


Figure 3 | TFIIE architecture and interactions. a, TFIIE interactions within the OC. Depicted are interactions of the TFIIE E-ribbon with the Pol II clamp, stalk subunit Rpb7, and the TFIIB B-ribbon, and interactions of the TFIIE eWH domain with the Pol II clamp helices and upstream DNA. The eWH E-wing lies close to the upstream DNA edge, similar to WH domains involved in DNA strand separation (Extended Data Fig. 3j). Colours as in Fig. 1, except for the Pol II stalk (Rpb4, dark red; Rpb7, dark blue). **b**, TFIIE domain architecture. The TFIIE variants used for functional assays are indicated as Cα spheres for point mutations, and with a black bracket for E-wing alterations (compare Extended Data Fig. 3g, h, j). Connectivity of the Tfa2 E-tether helices is uncertain. **c**, Selected TFIIE variants impair transcription from a *HIS4* promoter (Methods, Extended Data Fig. 3g, h). TFIIE-depleted nuclear extract (NE) was reconstituted with recombinant TFIIE or TFIIE variants carrying mutations in the Tfa1 eWH (M1, Tfa1(N50E/K51E/T52E); M2, Tfa1(N50A/K51A/T52A); M3, Tfa1(P56A/A59E/R62E); M4, Tfa1(ΔE-wing); M5, Tfa1(poly-Ala E-wing) and the Tfa1 E-ribbon (M6, Tfa1(L134E/V137E/L140E); M7, Tfa1(L134A/V137A/L140A) (Extended Data Fig. 3g). RNA products were visualized by primer extension and the mean intensity and standard deviation (s.d.) from triplicate experiments are provided, relative (rel.) to the activity of wild-type TFIIE. An asterisk marks RNA products resulting from an alternative upstream transcription start site.

and subunit dimerization (Extended Data Fig. 3a, refs 16, 38). The structure further indicates that TFIIE must be displaced or at least moved before the elongation factor Spt4/5 (human DSIF) can bind to polymerase[32,39].

The OC structure thus reveals how TFIIF and TFIIE bind open promoter DNA from opposite sides of the Pol II cleft. First, TFIIF adopts an extended induced structure that allows it to retain the upstream DNA–TBP–TFIIB assembly on the wall and to bind the DNA bubble and downstream duplex in the cleft (Figs 1b, 2). Second, the TFIIF Tfg2 WH domain and the TFIIE Tfa2 WH1 domain contact each other above upstream DNA to encircle and retain DNA. Third, the eWH domain of TFIIE binds DNA in the region of initial DNA opening and its long β1–β2 hairpin[33] (called here the 'E-wing') projects to the upstream edge of the bubble (Fig. 3a), suggesting that the eWH domain stabilizes open DNA. Taken together, the highly modular and flexible basal factors TFIIF and TFIIE undergo substantial induced folding transitions to engage in multiple protein–DNA and protein–protein interactions to stabilize the OC.

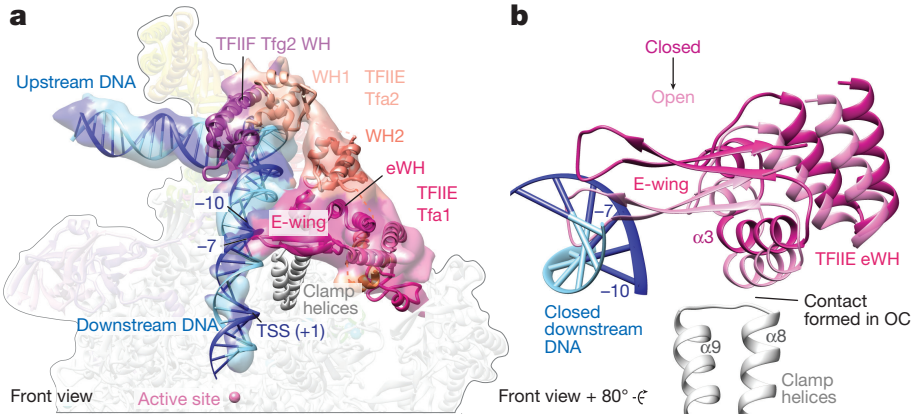


Figure 4 | Closed complex cryo-EM structure. a, Details of the closed complex viewed from the front[50]. Highlighted are TFIIE, TFIIF Tfg2 WH, and DNA, superimposed on their density. The promoter DNA displays increased flexibility downstream of the E-wing contact at position −7 upstream of the TSS (+1). **b**, Different positions of the TFIIE eWH in closed (dark magenta) and open (light magenta) complexes, viewed from the top[50]. Compare Extended Data Fig. 4e–g.

© 2016 Macmillan Publishers Limited. All rights reserved

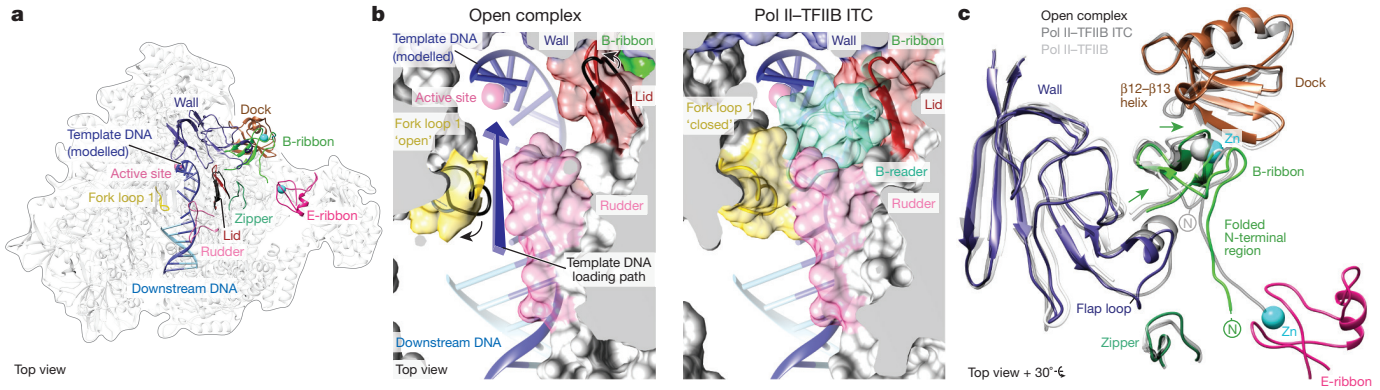


Figure 5 | Cleft clearance and DNA template loading. a, OC structure viewed from the top[50]. Highlighted are the Pol II active site (magenta sphere), fork loop 1 (yellow), lid (dark red), rudder (magenta), wall (navy blue), dock (brown), zipper (dark green), TFIIB B-ribbon (green), TFIIE E-ribbon (magenta) and downstream DNA (template, dark blue; non-template, cyan). The template single-strand was modelled using the Pol II–TFIIB ITC[15] crystal structure. **b**, Fork loop 1 and lid assume new positions in the OC compared to the ITC[15] and this opens a path (arrow) for loading of the template DNA strand (blue) into the active site (magenta sphere). Surface representations of Pol II cleft (silver), and cleft elements fork loop 1, lid, and rudder in the OC (left), and in a Pol II–TFIIB ITC[15] (PDB: 4BBS, right). Movement of the Pol II lid (left, black to dark red) leads to a steric clash with the B-reader (cyan). Compare Extended Data Fig. 5a–c. **c**, Allosteric binding of the TFIIE E-ribbon may lead to an altered position of the TFIIB B-ribbon. Movements in Pol II wall and flap loop (navy blue), dock (brown), zipper (dark green), and B-ribbon (green) are observed in presence of TFIIE compared to the crystal structures of the binary Pol II–TFIIB complex[13] (dark grey, PDB: 3K1F) and Pol II–TFIIB ITC[15] (light grey, PDB: 4BBS). The altered B-ribbon position may be stabilized by binding to a short helix formed in loop β12–β13 of the dock domain.

DNA opening and loading

Modelling of a closed DNA promoter onto the OC structure shows that closed DNA would clash with the TFIIE eWH domain. This suggests that the eWH domain adopts a different position before DNA opening. To investigate this, and to provide insights into the transition from the CC to the OC, we repeated structure determination with closed DNA instead of pre-opened DNA (Methods, Extended Data Fig. 4). Surprisingly, cryo-EM analysis revealed that about 3 out of 4 particles contained open DNA, although closed DNA was used for complex preparation. DNA opening occurred in the absence of TFIIH (Extended Data Fig. 4h, i). From these particles we obtained an independent reconstruction of the spontaneously formed OC at 6.1 Å resolution (OC5, Extended Data Fig. 4h). Weak density for upstream and downstream DNA segments indicates that DNA bubbles of various sizes formed during DNA opening (Extended Data Fig. 4i). The reconstruction resembles the high-resolution OC structure, suggesting that the latter was not perturbed by the use of pre-opened DNA (Extended Data Fig. 4h).

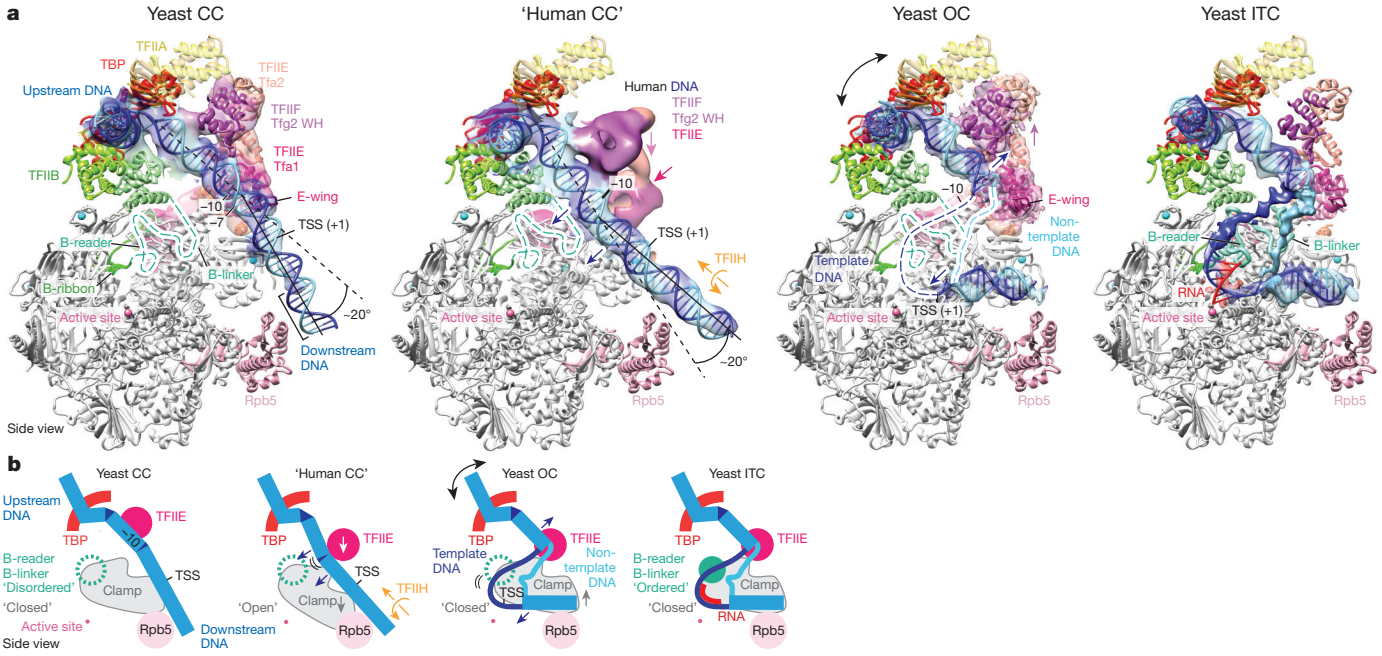


Figure 6 | Model for DNA opening during transcription initiation. **a**, Gallery of initiation complexes depicting proposed movements (arrows) of DNA and basal factors during the transition from the CC to OC to ITC, from left to right, viewed from the side[50]. Yeast CC and OC structures (this work) were complemented with our previous yeast ITC[19] structure (EMD-2785) and an alternative model of the CC ('human CC'), which was obtained by replacing the DNA with that in the human CC[20] (EMD-2306), and adjusting the clamp to the position observed in the human CC. Shown are cryo-EM densities for DNA, Tfg2 WH, and TFIIE. DNA positions −10, −7 (yeast CC) and +1 (TSS) are labelled. DNA was extended by one turn for the yeast CC (black bracket). The locations of TFIIA and TFIIE in the ITC were inferred from the yeast OC. Obstructing Pol II and TFIIF regions were removed for clarity. **b**, Schematic representation of **a**. Key elements for DNA opening are indicated (compare Extended Data Fig. 6).

© 2016 Macmillan Publishers Limited. All rights reserved

From the remaining particle images we obtained a reconstruction of the CC at 8.8 Å resolution (Fig. 4a). Comparison of this CC reconstruction with the OC structure reveals movements mainly in TFIIE (Fig. 4b, Extended Data Fig. 4e–g). In the CC, the E-wing lies on top of the DNA around position −7, in the region where DNA opening begins[7] (Extended Data Fig. 4e). Consistent with this contact, Tfa1 (or the human counterpart TFIIEα) crosslinks to DNA near this point in the CC[30,40,41] (Extended Data Fig. 4e). Conversion of the CC to the OC involves movement of upstream DNA and the DNA-associated domains Tfa2 WH1 and Tfg2 WH towards the cleft. DNA opening allows the TFIIE eWH domain to bind the tip of the clamp, and enables the E-wing to move near the upstream edge of the DNA bubble. Consistent with these changes, crosslinks between upstream DNA and the large TFIIE subunit are altered when the CC is converted to the OC[40]. These observations indicate that DNA opening involves TFIIE, and in particular the eWH domain.

TFIIE may also be involved in loading of the DNA template strand into the active centre during the transition from the CC to the OC. In previous structures, the path for loading the template was obstructed by the TFIIB B-reader and the Pol II fork loop 1 and lid. However, the B-reader is mobile in the OC structure, and fork loop 1 and the lid are moved to provide a path for template-strand loading (Fig. 5b, Extended Data Fig. 5a–c). These movements are apparently triggered by TFIIE because binding of the E-ribbon leads to a shift in the B-ribbon that partially withdraws the B-reader from the cleft (Fig. 5c). Thus allosteric binding of the E-ribbon apparently induces 'clearance' of the Pol II cleft that may facilitate template-strand loading into the active centre, and transcription start-site scanning in yeast.

To support the proposed functions of structural elements in TFIIE, we prepared recombinant TFIIE variants and tested them for binding to the CC and for promoter-dependent transcription activity in yeast nuclear extract (Methods, Fig. 3c, Extended Data Fig. 3g–i). Mutation of only three surface residues in the E-ribbon that contact the Pol II subunit Rpb7 strongly impaired both binding to the CC and transcription activity, and led to a severe growth defect in yeast (Fig. 3c, Extended Data Fig. 3h, j). Further, the eWH domain is required for TFIIE function[16], and disruption of the eWH contact with DNA by introducing negatively charged glutamate residues in the eWH helix α3 leads to a transcription defect and impairs binding to the CC (Fig. 3c, Extended Data Fig. 3h). Other mutations in the eWH domain did not show functional defects in our assays. Deletion of the E-wing results in a mild growth phenotype (Extended Data Fig. 3j), and does not impair *in vitro* transcription (Fig. 3c), maybe because TFIIH compensates for the loss of E-wing function in these assays.

Model of transcription initiation

Structural comparisons of the yeast CC, OC, and ITC[19] with each other and with the highly conserved human CC[20] reveal differences in the positions of DNA, the clamp, and TFIIE, and lead to an extended model of DNA opening (Fig. 6a, b, Extended Data Fig. 6). In this model, promoter DNA is initially bent away from the active site by ~20° near position −10 at the tip of the clamp helices (yeast CC). The clamp then opens slightly (Extended Data Fig. 6b), allowing promoter DNA to bend in the opposite direction and to enter the upper part of the cleft ('human CC'). This frees the site at the tip of the clamp helices that can now bind the TFIIE eWH domain. DNA can then no longer swing back to its initial position, because this would result in a steric clash with the repositioned eWH domain. However, when DNA opening occurs around position −10, the upstream DNA can swing back to its original position and this stabilizes the DNA duplex single-strand junction. TFIIH then rotates downstream DNA and pushes the template single strand into the cleft using its ATP-dependent translocase activity, as suggested previously[16]. When the bubble extends downstream[5], the template strand is loaded into the active centre cleft via a path cleared by allosteric binding of TFIIE (yeast OC). Template-strand loading allows the clamp to close again and to trap downstream DNA in the cleft (Extended Data Fig. 6b). The B-reader then covers the template strand, and helps to detect the transcription start site[13], triggering RNA synthesis (yeast ITC).

This model explains how DNA opening can be achieved with the use of binding energy alone, at least at some promoters[11,12]. DNA opening allows for new protein interactions at the upstream DNA duplex single-strand junction involving the eWH domain, the TFIIF arm, and the Pol II clamp. DNA loading into the cleft also enables new interactions of the downstream DNA duplex with the TFIIF charged helix and the Pol II cleft and clamp. These additional contacts may compensate for the energy needed for DNA melting and help to trap open DNA and to prevent its re-closure.

A similar mechanism for ATP-independent DNA opening may be used in other transcription systems. Proteins with homologies to TFIIB, TFIIE, and TFIIF, but not TFIIH, are present in the Pol I and Pol III systems[42], and counterparts of TBP, TFIIB, and TFIIE are found in archaea[38]. The bacterial transcription initiation system is structurally unrelated, but conceptually similar. DNA opening occurs spontaneously above the cleft, the open DNA is trapped with the use of binding energy, and this requires clamp opening and closure[43–47]. DNA opening at a subset of bacterial genes however requires the ATP-dependent initiation factor σ^{54}, and this enables further regulation[48]. Similarly, Pol II regulation can occur at the level of DNA opening[49], and TFIIH is required to keep DNA open at least at selected promoters[8]. This suggests that the Pol II initiation system evolved to depend on the ATP-consuming factor TFIIH for DNA opening, probably in response to an expanded need for gene regulation.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 17 November 2015; accepted 8 April 2016.
Published online 11 May 2016.

1. Buratowski, S., Hahn, S., Guarente, L. & Sharp, P. A. Five intermediate complexes in transcription initiation by RNA polymerase II. *Cell* **56,** 549–561 (1989).
2. Roeder, R. G. The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem. Sci.* **21,** 327–335 (1996).
3. Grünberg, S. & Hahn, S. Structural insights into transcription initiation by RNA polymerase II. *Trends Biochem. Sci.* **38,** 603–611 (2013).
4. Sainsbury, S., Bernecky, C. & Cramer, P. Structural basis of transcription initiation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.* **16,** 129–143 (2015).
5. Giardina, C. & Lis, J. T. DNA melting on yeast RNA polymerase II promoters. *Science* **261,** 759–762 (1993).
6. Ohkuma, Y. & Roeder, R. G. Regulation of TFIIH ATPase and kinase activities by TFIIE during active initiation complex formation. *Nature* **368,** 160–163 (1994).
7. Maxon, M. E., Goodrich, J. A. & Tjian, R. Transcription factor IIE binds preferentially to RNA polymerase IIa and recruits TFIIH: a model for promoter clearance. *Genes Dev.* **8,** 515–524 (1994).
8. Holstege, F. C., van der Vliet, P. C. & Timmers, H. T. Opening of an RNA polymerase II promoter occurs in two distinct steps and requires the basal transcription factors IIE and IIH. *EMBO J.* **15,** 1666–1677 (1996).
9. Parvin, J. D., Timmers, H. T. & Sharp, P. A. Promoter specificity of basal transcription factors. *Cell* **68,** 1135–1144 (1992).
10. Pan, G. & Greenblatt, J. Initiation of transcription by RNA polymerase II is limited by melting of the promoter DNA in the region immediately upstream of the initiation site. *J. Biol. Chem.* **269,** 30101–30104 (1994).
11. Buratowski, S., Sopta, M., Greenblatt, J. & Sharp, P. A. RNA polymerase II-associated proteins are required for a DNA conformation change in the transcription initiation complex. *Proc. Natl Acad. Sci. USA* **88,** 7509–7513 (1991).
12. Holstege, F. C., Tantin, D., Carey, M., van der Vliet, P. C. & Timmers, H. T. The requirement for the basal transcription factor IIE is determined by the helical stability of promoter DNA. *EMBO J.* **14,** 810–819 (1995).
13. Kostrewa, D. *et al.* RNA polymerase II-TFIIB structure and mechanism of transcription initiation. *Nature* **462,** 323–330 (2009).
14. Liu, X., Bushnell, D. A., Wang, D., Calero, G. & Kornberg, R. D. Structure of an RNA polymerase II-TFIIB complex and the transcription initiation mechanism. *Science* **327,** 206–209 (2010).
15. Sainsbury, S., Niesser, J. & Cramer, P. Structure and function of the initially transcribing RNA polymerase II–TFIIB complex. *Nature* **493,** 437–440 (2013).
16. Grünberg, S., Warfield, L. & Hahn, S. Architecture of the RNA polymerase II preinitiation complex and mechanism of ATP-dependent promoter opening. *Nat. Struct. Mol. Biol.* **19,** 788–796 (2012).

© 2016 Macmillan Publishers Limited. All rights reserved

17. Murakami, K. *et al.* Architecture of an RNA polymerase II transcription pre-initiation complex. *Science* **342,** 1238724 (2013).
18. Mühlbacher, W. *et al.* Conserved architecture of the core RNA polymerase II initiation complex. *Nat. Commun.* **5,** 4310 (2014).
19. Plaschka, C. *et al.* Architecture of the RNA polymerase II-Mediator core initiation complex. *Nature* **518,** 376–380 (2015).
20. He, Y., Fang, J., Taatjes, D. J. & Nogales, E. Structural visualization of key steps in human transcription initiation. *Nature* **495,** 481–486 (2013).
21. Murakami, K. *et al.* Structure of an RNA polymerase II preinitiation complex. *Proc. Natl Acad. of Sci. USA* **112,** 13543–13548 (2015).
22. Nikolov, D. B. *et al.* Crystal structure of a TFIIB–TBP–TATA-element ternary complex. *Nature* **377,** 119–128 (1995).
23. Geiger, J. H., Hahn, S., Lee, S. & Sigler, P. B. Crystal structure of the yeast TFIIA/TBP/DNA complex. *Science* **272,** 830–836 (1996).
24. Tan, S., Hunziker, Y., Sargent, D. F. & Richmond, T. J. Crystal structure of a yeast TFIIA/TBP/DNA complex. *Nature* **381,** 127–134 (1996).
25. Eichner, J., Chen, H.-T. T., Warfield, L. & Hahn, S. Position of the general transcription factor TFIIF within the RNA polymerase II transcription preinitiation complex. *EMBO J.* **29,** 706–716 (2010).
26. Deng, W. & Roberts, S. G. E. TFIIB and the regulation of transcription by RNA polymerase II. *Chromosoma* **116,** 417–429 (2007).
27. Fishburn, J. & Hahn, S. Architecture of the yeast RNA polymerase II open complex and regulation of activity by TFIIF. *Mol. Cell. Biol.* **32,** 12–25 (2012).
28. Čabart, P., Újvári, A., Pal, M. & Luse, D. S. Transcription factor TFIIF is not required for initiation by RNA polymerase II, but it is essential to stabilize transcription factor TFIIB in early elongation complexes. *Proc. Natl Acad. Sci. USA* **108,** 15786–15791 (2011).
29. Robert, F., Forget, D., Li, J., Greenblatt, J. & Coulombe, B. Localization of subunits of transcription factors IIE and IIF immediately upstream of the transcriptional initiation site of the adenovirus major late promoter. *J. Biol. Chem.* **271,** 8517–8520 (1996).
30. Forget, D., Langelier, M.-F. F., Thérien, C., Trinh, V. & Coulombe, B. Photo-cross-linking of a purified preinitiation complex reveals central roles for the RNA polymerase II mobile clamp and TFIIE in initiation mechanisms. *Mol. Cell. Biol.* **24,** 1122–1131 (2004).
31. Chen, H.-T. T., Warfield, L. & Hahn, S. The positions of TFIIF and TFIIE in the RNA polymerase II transcription preinitiation complex. *Nat. Struct. Mol. Biol.* **14,** 696–703 (2007).
32. Grohmann, D. *et al.* The initiation factor TFE and the elongation factor Spt4/5 compete for the RNAP clamp during transcription initiation and elongation. *Mol. Cell* **43,** 263–274 (2011).
33. Meinhart, A., Blobel, J. & Cramer, P. An extended winged helix domain in general transcription factor E/IIE alpha. *J. Biol. Chem.* **278,** 48267–48274 (2003).
34. Okuda, M. *et al.* A novel zinc finger structure in the large subunit of human general transcription factor TFIIE. *J. Biol. Chem.* **279,** 51395–51403 (2004).
35. Okuda, M. *et al.* Structural insight into the TFIIE-TFIIH interaction: TFIIE and p53 share the binding region on TFIIH. *EMBO J.* **27,** 1161–1171 (2008).
36. Okuda, M. *et al.* Structure of the central core domain of TFIIEβ with a novel double-stranded DNA-binding surface. *EMBO J.* **19,** 1346–1356 (2000).
37. Okamoto, T. *et al.* Analysis of the role of TFIIE in transcriptional regulation through structure-function studies of the TFIIEβ subunit. *J. Biol. Chem.* **273,** 19866–19876 (1998).
38. Blombach, F. *et al.* Archaeal TFEα/β is a hybrid of TFIIE and the RNA polymerase III subcomplex hRPC62/39. *eLife* **4,** e08378 (2015).
39. Martinez-Rucobo, F. W., Sainsbury, S., Cheung, A. C. M. & Cramer, P. Architecture of the RNA polymerase–Spt4/5 complex and basis of universal transcription processivity. *EMBO J.* **30,** 1302–1310 (2011).
40. Kim, T. K., Ebright, R. H. & Reinberg, D. Mechanism of ATP-dependent promoter melting by transcription factor IIH. *Science* **288,** 1418–1421 (2000).
41. Miller, G. & Hahn, S. A DNA-tethered cleavage probe reveals the path for promoter DNA in the yeast preinitiation complex. *Nat. Struct. Mol. Biol.* **13,** 603–610 (2006).
42. Vannini, A. & Cramer, P. Conservation between the RNA polymerase I, II, and III transcription initiation machineries. *Mol. Cell* **45,** 439–446 (2012).
43. Chakraborty, A. *et al.* Opening and closing of the bacterial RNA polymerase clamp. *Science* **337,** 591–595 (2012).
44. Feklistov, A. & Darst, S. A. Structural basis for promoter-10 element recognition by the bacterial RNA polymerase σ subunit. *Cell* **147,** 1257–1269 (2011).
45. Zhang, Y. *et al.* Structural basis of transcription initiation. *Science* **338,** 1076–1080 (2012).
46. Zuo, Y. & Steitz, T. A. Crystal structures of the *E. coli* transcription initiation complexes with a complete bubble. *Mol. Cell* **58,** 534–540 (2015).
47. Bae, B., Feklistov, A., Lass-Napiorkowska, A., Landick, R. & Darst, S. A. Structure of a bacterial RNA polymerase holoenzyme open promoter complex. *eLife* **4,** (2015).
48. Yang, Y. *et al.* Structures of the RNA polymerase-σ^{54} reveal new and conserved regulatory strategies. *Science* **349,** 882–885 (2015).
49. Kouzine, F. *et al.* Global regulation of promoter melting in naive lymphocytes. *Cell* **153,** 988–999 (2013).
50. Cramer, P., Bushnell, D. A. & Kornberg, R. D. Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science* **292,** 1863–1876 (2001).

Acknowledgements We thank C. Bernecky, W. Mühlbacher, S. Neyer, S. Sainsbury, and D. Tegunov for help and discussions; L. Larivière and L. Wenzeck for cloning and initial purification of TFIIE; W. Mühlbacher for initial cloning of TFIIA; S. Bilakovic for the modified pET-DUET-1 vector; J. Mahamid for help with data collection for the CC; K. Maier for help with yeast growth assays; M. Raabe and H. Urlaub for protein identification; K. Kinkelin for initial Pol II–TFIIF co-crystallization; and S. Hahn for providing the *TFA1* yeast strain and the shuffle plasmid pSH810. C.P. (SFB860), M.H. (GRK1721), and P.C. were supported by the Deutsche Forschungsgemeinschaft, the Advanced Grant TRANSIT of the European Research Council, and the Volkswagen Foundation.

Author Contributions C.P. designed and carried out high-resolution cryo-EM structure determinations of OC1–OC4. M.H. designed and carried out Pol II-TFIIF crystallographic analysis, and cryo-EM structure determinations of OC5 and CC. C.P. and M.H. designed and carried out functional assays. C.D. cloned and purified full-length TBP and TFIIA. C.D. and C.B. assisted with protein purification. J.P. supervised electron microscopy data collection. P.C. designed and supervised research. C.P., M.H., and P.C. prepared the manuscript.

Author Information Three-dimensional cryo-EM density maps of OC1, OC2, OC2-focused, OC3, OC3-focused, OC4, OC4-focused, OC5, and CC have been deposited in the Electron Microscopy Data Bank under the accession numbers EMD-3375, EMD-3376, EMD-3377, EMD-3378, EMD-3379, EMD-3380, EMD-3381, EMD-3382, and EMD-3383, respectively. Coordinate files of the OC and CC have been deposited in the Protein Data Bank under accession numbers 5FYW and 5FZ5. Coordinates and structure factors of the Pol II-Tfg1 peptide and the Pol II-TFIIF crystals have been deposited at the Protein Data Bank under the accession numbers 5IP7 and 5IP9. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to P.C. (patrick.cramer@mpibpc.mpg.de).

© 2016 Macmillan Publishers Limited. All rights reserved

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Vectors and sequences. The open reading frame (ORF) encoding full-length TBP was amplified from *Saccharomyces cerevisiae* (*Sc*) genomic DNA and cloned into a pOPINE vector, containing a C-terminal 6×histidine tag. Codon-optimized Tfg1, Toa1, and Toa2 were commercially obtained for expression in *Escherichia coli* (*E. coli*, Life Technologies). Toa1 (Δ95–209) was obtained by quick-change PCR and cloned with Toa2 into a pOPINE vector, containing a C-terminal 6×histidine-tag on Toa2. Tfa1 and Tfa2 ORFs were amplified from genomic *Sc* DNA and cloned sequentially into a pET vector. ORFs for Tfg1 and Tfg2 were cloned sequentially into a modified pET-Duet-1 vector, containing an N-terminal 10×histidine-8×arginine-SUMO-tag on Tfg1. Additional ribosomal binding sites were introduced as described[51]. For heterologous co-expression of *Sc* 16-subunit core Mediator (cMed)-Med1, 13-subunit cMed lacking Med4–Med9 was cloned with previously described vectors[19]. Three-subunit Med1–Med4–Med9 was prepared by cloning *Sc* ORFs of Med1, Med4, and Med9 into a modified pET-Duet-1 vector, containing an additional N-terminal 10×histidine-8×arginine-SUMO-tag on Med1. Sequences are available upon request.

Recombinant proteins. All proteins were expressed in *E. coli* BL21(DE3)RIL cells (Stratagene). The identity of all purified proteins was confirmed by mass spectrometry. All purified proteins and complexes were flash-frozen and stored at −80 °C.

Full-length TBP was expressed in *E. coli* that were grown in lysogeny broth (LB) medium at 37 °C to an optical density (OD) of 0.5 at 600 nm, and induced with 0.5 mM isopropyl-β-D-thiogalactoside (IPTG) for 4 h at 20 °C. Cells were lysed by sonication in buffer A (25 mM HEPES (pH 7.5), 500 mM KCl, 10% glycerol, 2.5 mM dithiothreitol (DTT)) containing protease inhibitors[52]. The soluble fraction was applied to a 5 ml Histrap HP column, washed with buffer A containing 1 M KCl, and eluted with buffer A containing 350 mM imidazole. The sample was diluted 1:4 with buffer A lacking KCl and applied to anion-exchange chromatography using a MonoS 5/50 (GE Healthcare), and eluted with a linear gradient of buffer A from 100–1000 mM KCl. TBP was further purified by size-exclusion chromatography using a Superose 12 10/300 (GE Healthcare) column, equilibrated in buffer B (10 mM HEPES (pH 7), 200 mM NaCl, 5% glycerol, 2 mM DTT). TBP-containing fractions were pooled and concentrated to 8 mg ml^{-1}.

TFIIA was obtained by co-expression of its subunits Toa1 and Toa2 in *E. coli*. Transformed cells were grown in LB medium at 37 °C to an OD of 0.5 at 600 nm. Expression was induced with 0.5 mM IPTG at 37 °C for 4 h. Cells were lysed in buffer C (25 mM Tris-HCl (pH 8.0), 500 mM NaCl, 10% glycerol, 2 mM DTT) containing protease inhibitors[52], and cleared by centrifugation. The supernatant was applied to a 5-ml HisTrap HP column, washed with buffer C containing 1 M NaCl, and eluted with buffer C containing 250 mM imidazole. Fractions containing the complex were pooled, diluted 1:4 with buffer C lacking NaCl, loaded on a MonoQ 5/50 anion-exchange column, and eluted with a linear gradient of buffer C from 100–500 mM NaCl. TFIIA containing fractions were pooled and applied to a Superdex 75 10/300 column (GE Healthcare), in buffer B. TFIIA was concentrated to 8 mg ml^{-1}. TFIIB was prepared as described[15].

Recombinant TFIIE was obtained by co-expression of its subunits Tfa1 and Tfa2 in *E. coli*. Cells were transformed and grown in LB medium at 37 °C to an OD of 0.6 at 600 nm, and expression was induced by 0.5 mM IPTG for 18 h at 18 °C. Cells were lysed by sonication in buffer D (50 mM Tris-HCl (pH 8.0), 300 mM NaCl, 0.02% Tween-20, 5 mM DTT) containing protease inhibitors[52]. The lysate was cleared by centrifugation and applied to a 5 ml Histrap HP column, equilibrated in buffer E (buffer M, lacking Tween-20, containing 10 mM imidazole). The column was washed with 10 column volumes of buffer E and eluted with buffer E containing 250 mM imidazole. TFIIE was then subjected to anion-exchange chromatography using a 5-ml Hi-Trap Heparin column (GE Healthcare), equilibrated in buffer F (50 mM Tris-HCl (pH 8.0), 100 mM NaCl, 2% glycerol, 5 mM DTT). The complex was eluted with a linear gradient of buffer F from 100–2000 mM NaCl. To improve purity, TFIIE was further applied to a Superose 12 10/300 size-exclusion column, in buffer G (5 mM HEPES (pH 7.25), 40 mM ammonium sulphate, 10 μM $ZnCl_2$, 10 mM DTT). TFIIE containing fractions were pooled and concentrated to 9.6 mg ml^{-1}.

Sc TFIIF subunits Tfg1 and Tfg2 were co-expressed in *E. coli* and cells were grown in LB medium at 37 °C to an OD of 0.8 at 600 nm. Expression was induced with 0.2 mM IPTG for 3 h at 37 °C. Cells were lysed by sonication in buffer H (50 mM HEPES (pH 7.0), 350 mM KCl, 10% glycerol, 2 mM DTT) supplemented with 50 mM imidazole and protease inhibitors[52]. Cleared lysate was applied to a 5-ml HisTrap HP column equilibrated in buffer H. The column was washed with 8 column volumes of buffer H containing 1000 mM KCl, and eluted with a linear gradient from buffer H to buffer I (50 mM HEPES (pH 7.0), 250 mM KCl, 800 mM imidazole, 10% glycerol, 2 mM DTT). The conductivity of the eluate was adjusted to match that of buffer J (50 mM HEPES (pH 7.0), 150 mM KCl, 10% glycerol, 2 mM DTT) and 3C protease cleavage was carried out for 2 h. The complex was then applied to cation-exchange chromatography using a 1-ml HiTrap SP HP column (GE Healthcare), equilibrated in buffer J, and eluted in a linear gradient from 150–1000 mM KCl. TFIIF was further purified by size-exclusion chromatography using a Superdex 200 10/300 Increase column (GE Healthcare), in buffer K (10 mM MES (pH 6.2), 150 mM KCl, 10% glycerol, 2 mM DTT). Purified TFIIF was concentrated to 3.9 mg ml^{-1}. For previous studies of the initiation complex we used the conserved *S. mikatae* Tfg1 (refs 18, 19), owing to difficulties in recombinant expression of its *Sc* homologue. Here, expression of *Sc* Tfg1 was enabled by codon optimization.

For preparation of recombinant 16-subunit cMed–Med1, 13-subunit cMed lacking Med4–Med9 was prepared essentially as described[19]. Separately, a plasmid containing Med1–Med4–Med9 was transformed in *E. coli* and cells were grown in LB medium at 37 °C to an OD of 0.6 at 600 nm. Protein expression was induced with 0.5 mM IPTG at 18 °C for 24 h. Cells were collected and lysed by sonication in buffer L (25 mM HEPES (pH 7.5), 400 mM potassium acetate, 10% glycerol, 20 mM imidazole (pH 8), 2 mM DTT) containing protease inhibitors[52]. Lysate was cleared by centrifugation and applied to a 5-ml Histrap HP column (GE Healthcare), equilibrated in buffer M (25 mM HEPES (pH 7.5), 400 mM potassium acetate, 10% glycerol, 30 mM imidazole (pH 8), 2 mM DTT). The column was washed with 5 column volumes of buffer M and a linear gradient over 10 column volumes from 30–100 mM imidazole. The heterotrimer was eluted with 300 mM imidazole over 10 column volumes. Fractions containing the complex were diluted 1:3 with buffer N (25 mM HEPES (pH 7.5), 100 mM KCl, 10% glycerol, 1 mM EDTA, 5 mM DTT) and incubated with 3C protease for 2 h on ice, to remove the affinity tag. The complex was further purified by anion-exchange chromatography using a MonoQ 5/50 GL column (GE Healthcare), equilibrated in buffer N, and was eluted with a linear gradient from 100–600 mM KCl over 150 column volumes. Fractions containing the complex were pooled and applied to size-exclusion chromatography using a Superose 6 10/600 column (GE Healthcare), equilibrated in buffer O (25 mM HEPES (pH 7.5), 200 mM KCl, 5 mM DTT). Purified Med1–Med4–Med9 complex was concentrated to 2.7 mg ml^{-1}. To prepare 16-subunit cMed–Med1, 13-subunit cMed was incubated with a twofold molar excess of Med1–Med4–Med9 for 30 min at 25 °C and 30 min on ice, and purified by size-exclusion using a Superose 6 10/600 column, in buffer P (25 mM HEPES (pH 7.5), 400 mM KCl, 5% glycerol, 5 mM DTT). Fractions containing cMed–Med1 were pooled and concentrated to 2 mg ml^{-1}.

Preparation of initiation complexes. Yeast 12-subunit Pol II was prepared as described[53]. Open and closed initiation complexes were prepared with several modifications of the previous cITC–cMed assembly scheme[19] and with a different nucleic acid scaffold for the closed complex. The 72 nucleotide nucleic acid scaffold previously used to prepare the Pol II core ITC (cITC)[19] contains a 15 nucleotide mismatch transcription bubble and six nucleotide RNA, and was used for assembly of the open initiation complex (OC). The closed complex (CC) contained a nucleic acid scaffold with 13 nucleotide longer downstream DNA, based on the *HIS4* promoter (template, 5′-TGATATTTTATGTATGTACA ACACACATCGGAGGTGAATCGAACGTTCCATAGCTATTATATACACAGCG TGCTACTGTTCTCG-3′; non-template, 5′-CGAGAACAGTAGCACGCTGTGT ATATAATAGCTATGGAACGTTCGATTCACCTCCGATGTGTGTTGTACATA CATAAAAATATCA-3′). The initiation complex was prepared as follows. Pol II (200 μg at 3 mg ml^{-1}) was incubated with a fourfold molar excess of TFIIF. A twofold molar excess of nucleic acid scaffold over Pol II, tenfold molar excess of TFIIA, fourfold molar excess of TBP and TFIIB were added to buffer Q (25 mM HEPES-KOH (pH 7.5), 150 mM potassium acetate, 5% glycerol, 2 mM $MgCl_2$, 5 mM DTT) and incubated with pre-formed Pol II–TFIIF complex for 8 min at 25 °C. TFIIE was added in a tenfold molar excess over Pol II and incubated for 5 min at 25 °C. cMed–Med1 was added in a 1.2-fold molar excess over Pol II and incubated for 50 min at 25 °C. Open and closed complexes were purified using a Superose 6 3.2/300 size exclusion column (GE Healthcare), equilibrated in buffer Q. Fractions containing the complex were pooled (0.4–0.8 mg ml^{-1}) and additionally incubated with equimolar amounts of nucleic acid scaffold. The sample was then crosslinked for 30 min on ice using 0.1% glutaraldehyde (Electron Microscopy Sciences), and the reaction was quenched with 50 mM lysine (Sigma). The crosslinked sample was re-purified in a second size-exclusion step using a Superose 6 3.2/300 column, equilibrated in buffer Q lacking glycerol. Fractions containing initiation complexes were pooled (0.2–0.6 mg ml^{-1}) and used for EM grid preparation.

Electron microscopy. Initiation complex samples were applied to R3.5/1 holey carbon grids (Quantifoil). Grids were glow-discharged for 15 s before deposition of 4.5 μl complex, and subsequently blotted and vitrified by plunging into

© 2016 Macmillan Publishers Limited. All rights reserved

liquid ethane with a Vitrobot Mark IV (FEI) operated at 4 °C and 100% humidity. Cryo-EM data was acquired on a FEI Titan Krios operated in EFTEM mode at 300 keV, and equipped with a K2 Summit direct detector (Gatan). Automated data collection was carried out using the TOM toolbox[54] to acquire 1756 movies of the OC with a range of defocus values (from −0.7 μm to −4.2 μm) at a nominal magnification of 37,000× (1.35 Å per pixel). The camera was operated in 'super-resolution' mode (0.675 Å per pixel), with a total exposure time of 10 s fractionated into 25 frames, a dose rate of ∼5 e^- per pixel per second, and total dose of 33 e^- $Å^{-2}$ per movie. Cryo-EM data of the CC was collected in the same manner, except that 959 movies were acquired with a defocus range from −0.8 μm to −5 μm, a total exposure time of 6 s with 20 frames, a dose rate of ∼8 e^- per pixel per second, and total dose of 40 e^- $Å^{-2}$. Movies were aligned as described[19,55], except that images were not partitioned into quadrants.

Image processing. For single particle analysis of the OC, an initial set of 10,225 particles was selected semi-automatically using e2boxer.py from EMAN2 (ref. 56). CTF parameters were estimated using CTFFIND4 (ref. 57). CTF correction and subsequent image processing was performed with RELION 1.3 (ref. 58), unless otherwise noted. Resolution was reported on the basis of the gold-standard Fourier shell correlation (FSC) (0.143 criterion) as described previously[59] and temperature factors were determined and applied automatically in RELION[58]. Selected particles were extracted with a 300^2 pixel box and pre-processed to normalize images. Reference-free two-dimensional (2D) class averages were calculated, and twelve representative classes were low-pass filtered to 25 Å resolution and used as templates for automated picking[60] of all micrographs. The resulting 415,030 particle images were screened manually and by reference-free 2D classification, yielding 257,259 particle images that were used for subsequent processing. The 7.8 Å cryo-EM map of the yeast cITC[19] (EMD-2785) was low-pass filtered to 50 Å and used as initial model for 3D refinement of the 10,225 particle set. This revealed an OC density to an estimated resolution of 10 Å. This density was low-pass filtered to 50 Å and used for processing of the complete OC cryo-EM data set (Extended Data Fig. 7a–d). A 3D reconstruction of all particles was calculated to 3.75 Å, and subjected to particle polishing using RELION 1.4 beta[58]. This lead to an improved density at 3.58 Å resolution.

Hierarchical 3D classification was carried out without image alignment, to reduce computational requirements and identify homogeneous single particle groups (Extended Data Fig. 7a–d). Soft masks encompassing the complete OC or smaller regions of Pol II and basal factors were generated using the volume eraser in UCSF Chimera[61] and RELION[58]. This included masks for Pol II stalk–TFIIE E-ribbon (Extended Data Fig. 7a, c), upstream DNA–TBP–TFIIA–TFIIB–Tfg2 linker (Extended Data Fig. 7b), TFIIE–Tfg2 WH (Extended Data Fig. 7c), TFIIF (Extended Data Fig. 7d), and TFIIF dimerization domain (Extended Data Fig. 7d). Each class was refined using the 3D auto-refine procedure against the respective particles within that class with a soft reference mask in the shape of the OC (maximum diameter of 270 Å), generated in RELION[58]. The OC1 reconstruction (improved Pol II core, B-ribbon, and E-ribbon density) was determined from 102,876 particles to a resolution of 3.58 Å and with a temperature factor of −111 $Å^2$ (Extended Data Fig. 8a). The OC2 reconstruction (improved upstream DNA, TBP, TFIIA, TFIIB, and Tfg2 linker density) was determined from 17,282 particles to a resolution of 3.97 Å and with a temperature factor of −95 $Å^2$ (Extended Data Fig. 8b). The OC3 reconstruction (improved TFIIE density) was determined from 11,231 particles to a resolution of 4.35 Å and with a temperature factor of −125 $Å^2$ (Extended Data Fig. 8c). The OC4 reconstruction (improved TFIIF dimerization and Tfg1 arm density) was determined from 29,455 particles to a resolution of 3.89 Å and with a temperature factor of −79 $Å^2$ (Extended Data Fig. 8d). Focused refinement of the upstream DNA assembly (OC2-focused) was achieved by continuing the auto-refinement (round 1, class 4; Extended Data Fig. 7c) from the first round of local searches using a corresponding soft mask. Focused refinement of TFIIE–Tfg2 WH (OC3-focused) was achieved by continuing the OC3 auto-refinement from the first round of local searches using a soft mask encompassing TFIIE, Tfg2 WH, and the interacting segment of upstream DNA. Focused refinement of TFIIF dimerization domain (OC4-focused) was achieved by continuing the auto-refinement (round 2, class 2; Extended Data Fig. 7d) from the first round of local searches using a soft mask encompassing the TFIIF dimerization domain and the Pol II lobe. The resolution of focused refinements was determined using a soft mask with a 30-pixel soft edge[62], to 4.7 Å, 7.5 Å, and 4.09 Å for OC2-, OC3-, and OC4-focused refinements respectively (Extended Data Fig. 8b, c, d).

Single particle cryo-EM analysis of the closed complex was carried out essentially as for the open complex, with the following differences. Reference-free 2D classification of 10,591 particles, picked semi-automatically using e2boxer.py from EMAN2[56], gave three representative 2D class averages that were low-pass filtered to 30 Å resolution and used as templates for automated particle picking in RELION of all micrographs[60]. The resultant 155,079 particles were screened manually and by reference-free 2D classification, yielding 111,625 particles that were used for subsequent processing. The cITC (EMD-2785), low-pass filtered to 50 Å, was used as the reference model for the initial 10,591 particle set, and the resultant 3D reconstruction (14 Å resolution), again low-pass filtered to 50 Å, was used as initial reference model for 3D refinement using the full cryo-EM data set. This revealed the CC at 7.5 Å, and was improved by particle polishing to 6.5 Å resolution using RELION 1.4 beta[58]. Hierarchical 3D classification without image alignment using a soft mask encompassing the complete CC resulted in two populations, CC and OC5, respectively (Extended Data Fig. 7e). Both classes were refined using the 3D auto-refine procedure against the respective particles within that class with a soft mask in the shape of the CC. The CC soft mask was also suitable for the OC as both complexes have a similar shape. The CC reconstruction was determined from 7,527 particles to a resolution of 8.2 Å. The OC5 reconstruction was determined from 79,797 particles to a nominal resolution of 6.1 Å and a temperature factor of −176 $Å^2$ was applied (Extended Data Fig. 8g). To improve the densities TFIIE, Tfg2 WH and downstream DNA in the CC, we carried out focused classification of the Pol II stalk and subsequently TFIIE, Tfg2 WH and downstream DNA using soft masks encompassing these regions. The individual classes were refined as before and yielded a CC at 8.8 Å comprising 5,690 particles and a temperature factor of −300 $Å^2$ was applied (Extended Data Fig. 8f). A similar focused classification scheme for OC5 revealed moderately improved density for TFIIE (not shown).

Local resolution estimates were determined using a sliding window of 40^3 voxels as previously described except that a single pair two half-maps was used and resolution estimates were not capped at the nominal resolution[19] as no local filters were applied.

Structural modelling. A composite model of the OC was obtained using cryo-EM densities OC1–OC4 and the focused refinements of OC2, OC3, and OC4 particles. Structural models were built in COOT[63] unless indicated otherwise. Models were refined using the real space refinement routine in Phenix[64] into the respective OC density, as indicated, with secondary structure and rotamer restraints. First, structural models of Pol II (lacking Rpb4–Rpb7), TFIIB B-ribbon (residues 22–59), and downstream DNA (PDB: 4V1N[19]) were placed into the OC1 map using UCSF Chimera[61], followed by rigid-body group refinement in Phenix[64]. Models of the Pol II core and TFIIB B-ribbon were adjusted and extended manually, and refined in Phenix[64] into the OC1 density. The Rpb4–Rpb7 structure (PDB: 4V1N) was fitted, and residues at the base of the stalk (Rpb7 1–81, 150–159) were modified to fit the density and refined into the OC1 map. TFIIE was modelled using structural information on eWH[33], E-ribbon[34], and WH1[36], and a homology model for WH2 (Fig. 3a, b). A homology model of the human TFIIEα E-ribbon (PDB: 1VD4 (ref. 34)) was generated using MODELLER[65] (Tfa1 residues 121–158) and rigid-body fitted into the unsharpened OC1 density. Upstream DNA (PDB: 4V1N), TFIIB N-terminal cyclin (PDB: 4BBR, chain M and residues 122–213), the homology model of TFIIB C-terminal cyclin domain (PDB: 4V1N, chain M and residues 233–343) were fitted into the OC2 cryo-EM density. Models of Pol II protrusion, wall and the TFIIB N-terminal cyclin were adjusted and extended manually, and the TFIIB C-terminal cyclin (residues 236–328) and Tfg2 linker regions (residues 249–280) were built into OC2 and OC2-focused maps and refined into the OC2 map in Phenix[64]. X-ray structures of TBP (PDB: 1YTB[66], chain A), and TFIIA (PDB: 1YTF[25], chains B–D) were individually fitted into the OC2 map obtained from focused refinement (OC2-focused). The fit of upstream DNA, from the upstream edge of the DNA bubble to TBP, was improved in UCSF Chimera[61] using the OC3 map. Protein models of Tfa1 eWH (residues 1–90) and Tfa2 WH2 (residues 187–249), generated with the I-TASSER prediction server[67], were fit into the OC3 density and adjusted. The TFIIE linker helices were built *de novo* and modified to match the density, and together with Tfa1 eWH and Tfa2 WH2 models were refined into the unsharpened OC3 map in Phenix[64] and subjected to the phenix.model_idealization routine to optimize geometry. Published NMR data of the Tfg2 WH domain[68] (BMRB accession number 17916) was used to calculate a low-energy model with the BMRB CS-ROSETTA server[69]. The Tfg2 WH and an I-TASSER model of Tfa1 WH1 (residues 123–181) were placed into the OC3 density obtained from focused refinement (OC3-focused). The Tfg2 WH positioning is also consistent with NMR data for the Tfg2 WH–DNA interface[68]. A part of the Tfg1 arm domain (residues 327–349) was built into OC3 and OC4 densities and refined into the OC4 density. A previous homology model of the TFIIF dimerization domain[18] was used as a starting template for modelling into OC4 and OC4-focused density maps, and was refined into the OC4 density. The crystal structure of the Tfg1 N-terminal peptide (residues 21–35, see below) was fitted into an OC4 density (round 2 class 2, Extended Data Fig. 7d). The individually refined models showed good stereochemistry and were validated with Molprobity[70], and the FSC of map versus model (Extended Data Figs 8 and 9).

To generate the CC model, the OC model of Pol II, TFIIA, TFIIB, TBP, TFIIE E-ribbon and TFIIF dimerization domain was rigid-body fitted into the CC density

© 2016 Macmillan Publishers Limited. All rights reserved

using an automated global 6D correlation search in Situs[71]. The remaining protein part was divided into two rigid bodies containing (i) Tfa1 and Tfa2 WH2, and (ii) Tfa2 WH1 and Tfg2 WH, which were independently fitted into the CC density using Situs[71]. To model closed promoter DNA, upstream DNA was extended with canonical duplex B-form DNA in COOT[63] and rigid-body fitted in UCSF Chimera[57] to reflect the density.

All figures were generated using UCSF Chimera[61].

Crystallographic analysis of Pol II–TFIIF complex. The structure of the Tfg1 N-terminal region (residues 21–35) bound to Pol II was determined by X-ray crystallographic analysis. 12-subunit Pol II was prepared and crystallized[53], and TFIIF or selenomethionine (SeMet)-substituted Tfg1 peptide (residues 19–41, Peptide Speciality Laboratories GmBH, Heidelberg) was added to the cryo-protectant as described[72]. Diffraction data were collected on a PILATUS 6M detector at the X06SA beamline (SLS, Villigen, Switzerland) at 100 K. Data for the Pol II–Tfg1 SeMet peptide crystal was collected at a wavelength of 0.97972 Å and for the Pol II–TFIIF crystal at 0.91889 Å. Data were processed with XDS and XSCALE[73]. The structure was phased with the crystal structure of the 12-subunit Pol II (PDB: 3PO2) lacking nucleic acid and refined using BUSTER[74]. Model building of the TFIIF N-terminal residues in COOT[63] was guided by a selenium anomalous difference Fourier peak (M27). Subsequent refinement in BUSTER used secondary structure restraints for the Tfg1 peptide helix. The final structures had an R_{free} factor of 18.0% and 19.4% for the Pol II–Tfg1 SeMet peptide crystal and the Pol II–TFIIF crystal, respectively (Extended Data Fig. 9c), and showed good stereochemistry[70]. In the Pol II–Tfg1 SeMet peptide structure 90% of the residues fall in favoured regions of the Ramachandran plot and 3% in disallowed regions. For the Pol II–TFIIF structure 89% of the residues fall in favoured regions of the Ramachandran plot and 3% in disallowed regions.

Yeast and functional assays. TFIIE interfaces with Pol II and DNA were probed by *in vivo* mutagenesis. The yeast strain used for *TFA1* genetic assays[16] was provided by S. Hahn (Fred Hutchinson Cancer Research Center). Mutations were introduced into a plasmid encoding *TFA1* (pSH810, ARS CEN LEU2 3×Flag), as indicated in Extended Data Fig. 3g. The *tfa1*(ΔE-wing) mutant was generated by deletion of *TFA1* residues 71–84, and their replacement by the residues GSG. The *TFA1* constructs were transformed into the shuffle strain, and were streaked twice on −Ura −Leu plates, and subsequently onto yeast extract-peptone-dextrose (YEPD) plates. Yeast were freshly grown in YEPD medium and resuspended in water to an OD of 1 at 600 nm, and tenfold dilutions were spotted on 5-fluoroorotic acid (5-FOA) and YEPD plates, and incubated at 30 °C.

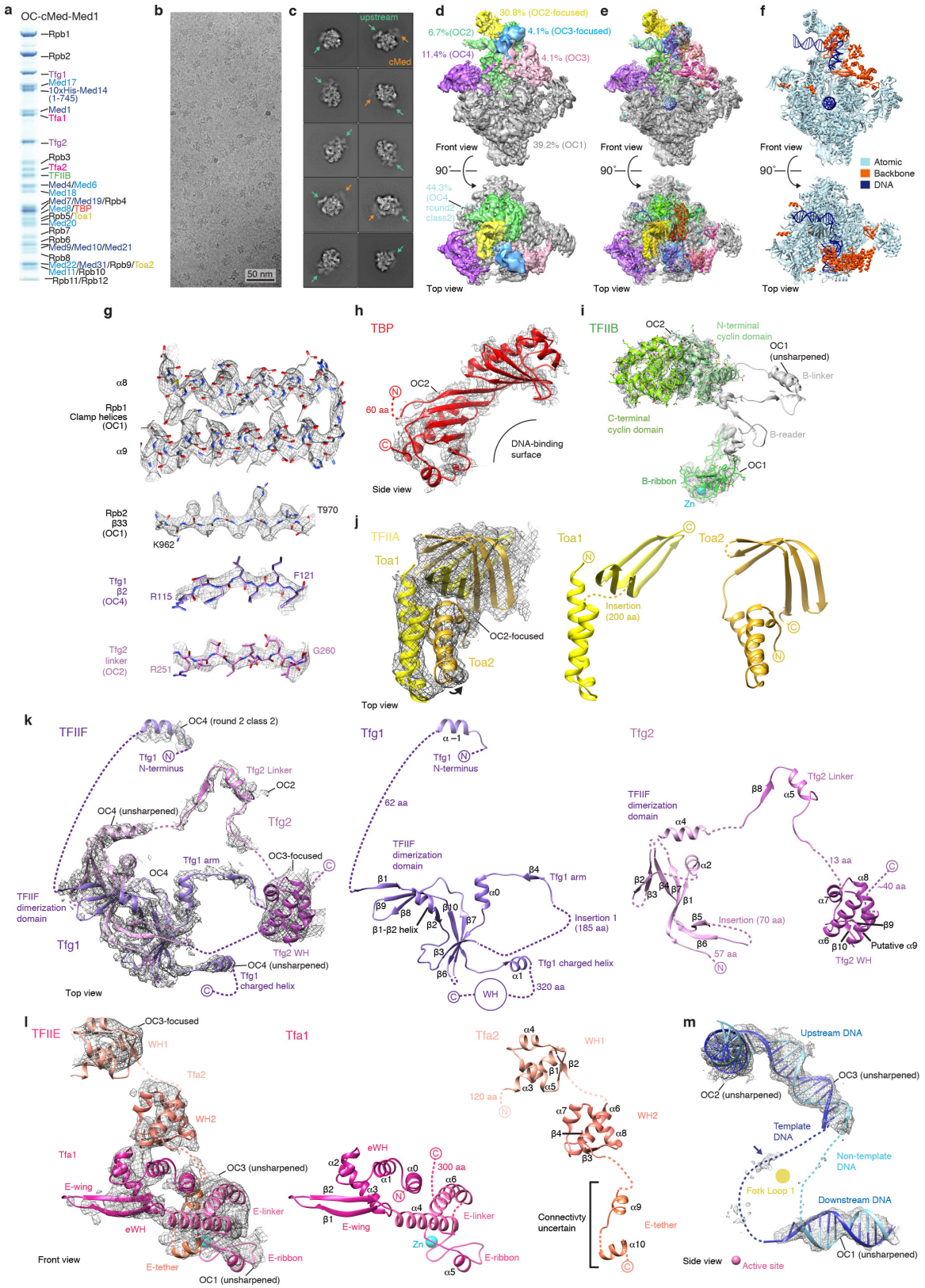
To further characterize the interfaces between TFIIE and the initiation complex, recombinant TFIIE mutants were purified according to the same protocol as wild-type TFIIE (Recombinant proteins, Extended Data Fig. 3g, h). The Tfa1(poly-Ala E-wing) mutant was generated by replacement of *TFA1* residues 71–84 with poly-Alanine.

To assess the interaction of wild-type TFIIE and the TFIIE mutants with the CC by protein pulldown, 3 μg purified Pol II was first biotinylated on the Rpb3 subunit as described[19]. The CC, containing biotinylated Pol II, was subsequently prepared as above ('preparation of initiation complexes'), but without TFIIE. This preparation was immobilized on 15 μL Dynabeads M280 streptavidin resin (Life Technologies), equilibrated in buffer Q. 10 μg TFIIE or TFIIE mutant (tenfold molar excess over Pol II) was incubated with the immobilized CC or control beads for 1 h at 4 °C. The beads were washed four times and bound proteins were analysed by SDS–PAGE (Extended Data Fig. 3h).

Promoter-dependent *in vitro* transcription was used to determine the activity of TFIIE and the TFIIE mutants. The yeast strain used for *TFA1* genetic assays[16] was transformed with a plasmid containing 3×Flag-tagged *TFA1* (pSH810, ARS CEN LEU2 3×Flag). Transformants were streaked once onto −Ura −Leu plates, once onto −Leu plates, twice onto 5-FOA plates, and subsequently onto YEPD plates. Nuclear extract from 3 l yeast culture was prepared as described[19]. The nuclear extract was immunodepleted of 3×Flag–Tfa1 as described[19] with the following modifications. Before nuclear extract was incubated with anti-Flag M2 agarose beads (Sigma), beads were incubated with 1 mg ml^{-1} bovine serum albumin (BSA) protein (Sigma) for 1 h at 4 °C on a turning wheel followed by three wash steps. Immunodepleted nuclear extract was separated from beads by Micro Bio-Spin chromatography columns (Biorad). Specificity of the depletion was confirmed by western blot carried out as described previously[19] for 3×Flag-tagged Tfa1 (Macs Miltenyi Biotec, 130-101-572), Rpb3 (Neoclone, WP012), TFIIB (Abcam, sc-274) and Histone H3 (Abcam, ab21054) (Extended Data Fig. 3i). The secondary antibodies anti-rabbit IgG horse-radish peroxidase (HRP; GE Healthcare, NA934) and anti-mouse IgG HRP (Abcam, ab5870) were used (Extended Data Fig. 3i). Activator- and promoter-dependent *in vitro* transcription and primer extension were carried out as described previously[19]. Recombinant TFIIE (5 pmol) and TFIIE mutants (5 pmol) were added to the depleted nuclear extract as indicated in Fig. 3c. Transcripts were visualized on a denaturing 8% polyacrylamide TBE gel with a Typhoon 9500 scanner (GE Healthcare) and quantified with ImageQuant (GE Healthcare). For quantification, the relative activity of each variant compared to TFIIE was calculated for each replicate. The mean intensity and standard deviation of three replicates was calculated from their relative activities. A second RNA product from the HIS4 promoter was observed, apparently resulting from an alternative upstream transcription start site. Although some differences in the relative use of the two TSSs are observed in the assay, we refrained from interpreting these.

51. Baumli, S., Hoeppner, S. & Cramer, P. A conserved mediator hinge revealed in the structure of the MED7.MED21 (Med7.Srb7) heterodimer. *J. Biol. Chem.* **280,** 18171–18178 (2005).
52. Seizl, M. *et al.* A conserved GA element in TATA-less RNA polymerase II promoters. *PLoS One* **6,** e27595 (2011).
53. Sydow, J. F. *et al.* Structural basis of transcription: mismatch-specific fidelity mechanisms and paused RNA polymerase II with frayed RNA. *Mol. Cell* **34,** 710–721 (2009).
54. Korinek, A., Beck, F., Baumeister, W., Nickell, S. & Plitzko, J. M. Computer controlled cryo-electron microscopy – TOM^2 a software package for high-throughput applications. *J. Struct. Biol.* **175,** 394–405 (2011).
55. Li, X., Zheng, S. Q., Egami, K., Agard, D. A. & Cheng, Y. Influence of electron dose rate on electron counting images recorded with the K2 camera. *J. Struct. Biol.* **184,** 251–260 (2013).
56. Tang, G. *et al.* EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157,** 38–46 (2007).
57. Rohou, A. & Grigorieff, N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192,** 216–221 (2015).
58. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180,** 519–530 (2012).
59. Chen, S. *et al.* High-resolution noise substitution to measure overfitting and validate resolution in 3D structure determination by single particle electron cryomicroscopy. *Ultramicroscopy* **135,** 24–35 (2013).
60. Scheres, S. H. Semi-automated selection of cryo-EM particles in RELION-1.3. *J. Struct. Biol.* **189,** 114–122 (2015).
61. Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).
62. Nguyen, T. H. *et al.* The architecture of the spliceosomal U4/U6.U5 tri-snRNP. *Nature* **523,** 47–52 (2015).
63. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60,** 2126–2132 (2004).
64. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66,** 213–221 (2010).
65. Eswar, N. *et al.* Comparative protein structure modeling using Modeller. *Current Protoc. Bioinformatics* http://dx.doi.org/10.1002/0471250953.bi0506s15 (2006).
66. Kim, Y., Geiger, J. H., Hahn, S. & Sigler, P. B. Crystal structure of a yeast TBP/TATA-box complex. *Nature* **365,** 512–520 (1993).
67. Yang, J. *et al.* The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* **12,** 7–8 (2015).
68. Kilpatrick, A. M., Koharudin, L. M., Calero, G. A. & Gronenborn, A. M. Structural and binding studies of the C-terminal domains of yeast TFIIF subunits Tfg1 and Tfg2. *Proteins* **80,** 519–529 (2012).
69. Shen, Y. *et al.* Consistent blind protein structure generation from NMR chemical shift data. *Proc. Natl Acad. Sci. USA* **105,** 4685–4690 (2008).
70. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66,** 12–21 (2010).
71. Wriggers, W. Using Situs for the integration of multi-resolution structures. *Biophys. Rev.* **2,** 21–27 (2010).
72. Kinkelin, K. *et al.* Structures of RNA polymerase II complexes with Bye1, a chromatin-binding PHF3/DIDO homologue. *Proc. Natl Acad. Sci. USA* **110,** 15277–15282 (2013).
73. Kabsch, W. XDS. *Acta Crystallogr. D* **66,** 125–132 (2010).
74. Blanc, E. *et al.* Refinement of severely incomplete structures with maximum likelihood in BUSTER-TNT. *Acta Crystallogr. D* **60,** 2210–2221 (2004).
75. Gaiser, F., Tan, S. & Richmond, T. J. Novel dimerization fold of RAP30/RAP74 in human TFIIF at 1.7 A resolution. *J. Mol. Biol.* **302,** 1119–1127 (2000).
76. Ha, I. *et al.* Multiple functional domains of human transcription factor IIB: distinct interactions with two general transcription factors and RNA polymerase II. *Genes Dev.* **7,** 1021–1032 (1993).
77. Harami, G. M., Gyimesi, M. & Kovács, M. From keys to bulldozers: expanding roles for winged helix domains in nucleic-acid-binding proteins. *Trends Biochem. Sci.* **38,** 364–371 (2013).
78. Esnault, C. *et al.* Mediator-dependent recruitment of TFIIH modules in preinitiation complex. *Mol. Cell* **31,** 337–346 (2008).
79. Čabart, P. & Luse, D. S. Inactivated RNA polymerase II open complexes can be reactivated with TFIIE. *J. Biol. Chem.* **287,** 961–967 (2012).

© 2016 Macmillan Publishers Limited. All rights reserved

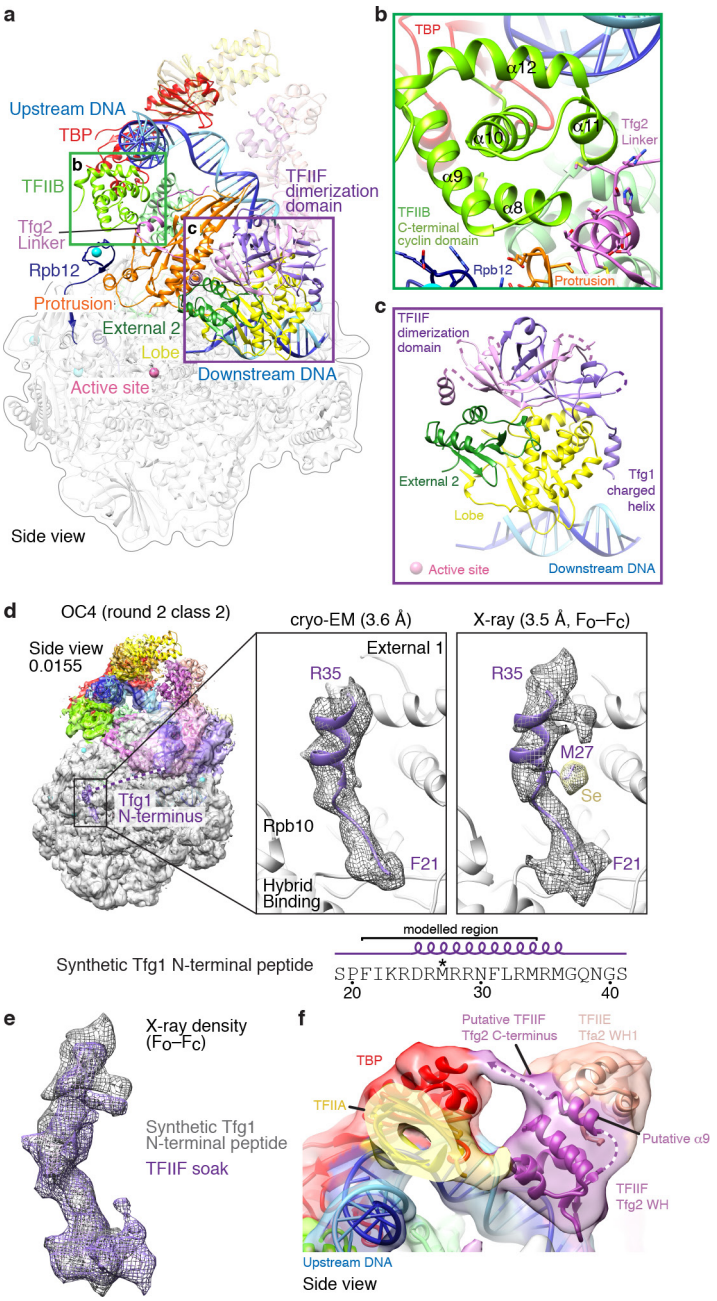


Extended Data Figure 1 | See next page for caption.

© 2016 Macmillan Publishers Limited. All rights reserved

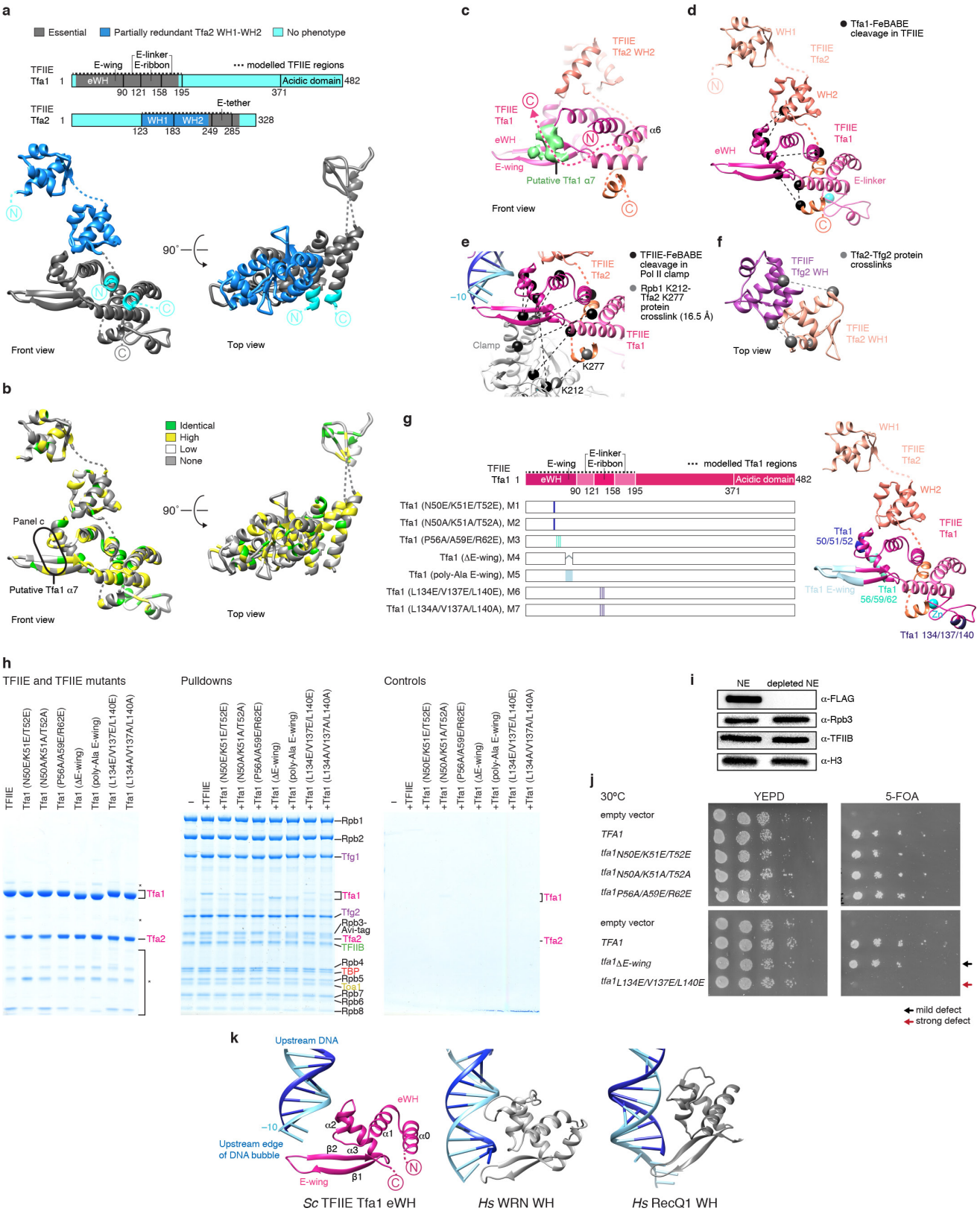
Extended Data Figure 1 | Modelling of open complex cryo-EM densities. a, SDS–PAGE analysis of OC–cMed–Med1 complex after size-exclusion chromatography. Protein colours as in Fig. 1. Although core Mediator was required for stable association of TFIIE, it largely dissociated under cryo-EM conditions as observed previously[19]. Some remaining core Mediator was flexible and located as described previously[19], but could not be included in further high-resolution analysis. **b**, Cryo-EM micrograph of the OC–cMed–Med1 complex. Scale bar, 50 nm. **c**, Ten representative reference-free 2D class averages of OC–cMed–Med1 reveal flexibility of the upstream DNA assembly including TFIIE (green arrow) and very weak density for core Mediator (orange arrow). Compare Extended Data Fig. 7a, c. **d**, Composite cryo-EM density of the OC shown in front and top views[50]. Colours indicate the cryo-EM densities used for modelling of the open complex (OC1, grey; OC2, green; OC2-focused, yellow; OC3, salmon; OC3-focused, blue; OC4 purple; OC4 round 2 class 2, light blue). Shown are the unsharpened cryo-EM densities. The percentage of particles from the full set of 257,259 that was used for the respective reconstruction is indicated. **e**, Composite cryo-EM density of the OC superimposed on a ribbon model of the OC, coloured as in Fig. 1. The composite cryo-EM density enabled modelling of the initiation factors and DNA. Our structure also enabled correction of the revised yeast initiation complex model obtained by Murakami *et al.* from cryo-EM at 6 Å resolution[21], and we note the following differences between the structures, superimposed on Rpb1: (1) The TFIIF Tfg2 WH domain is rotated by ~180°, which is further inconsistent with nuclear magnetic resonance (NMR) data on the TFg2 WH–DNA interface[68] and fits comparatively worse to protein–protein crosslinking data between the Tfg2 WH and Tfa2 WH1 (ref. 17). (2) Domains of TFIIE, except Tfa2 WH1, were placed incorrectly: Tfa1 eWH (rotation and translation into the E-linker density; 17 Å distance for helix α3 in our CC), Tfa1 E-ribbon (rotation and translation into E-linker density; 35 Å distance between the Zn atoms), and Tfa2 WH2 (~180° rotation). Further, the Tfa2 E-tether region was incorrectly assigned to density belonging to the Tfa1 eWH. The Tfa1 E-linker was not modelled. (3) The TFIIF Tfg1 arm was modelled into an empty space lacking density, and the Tfg1 helix α0 was absent. Our models of the TFIIF dimerization domain, Tfg2 linker, Tfg1 N terminus, and Tfg1 arm fit into densities from a recent study[21], indicating the electron microscopic reconstruction is correct, but that the modelling was premature at the available resolution. **f**, Ribbon model of the OC coloured according to how different parts of the OC were modelled into the OC cryo-EM densities (see **d**). Regions with atomic (light blue) and backbone models (orange), and DNA (dark blue) are indicated. Views as in **d**. **g**, Representative regions of the sharpened cryo-EM densities OC1 (3.6 Å), OC2 (4.0 Å), and OC4 (3.9 Å) are shown with the underlying refined coordinate model. The OC1 density shows clear side-chain features for Rpb1 clamp helices α8 and α9 and Rpb2 β33, the OC4 density for Tfg1 β2 that is part of the dimerization domain, and the OC2 density for part of the Tfg2 linker. For OC nomenclature, see Extended Data Fig. 7. **h**, Fit of the TBP crystal structure (PDB: 1YTB)[66] to the OC2 cryo-EM density, shown in a Pol II side view[50]. **i**, Fit of TFIIB N- and C-terminal cyclin domains, B-linker and B-reader, and B-ribbon elements to OC1 and OC2 cryo-EM densities. The B-linker element displays weak density, and the B-reader is not observed. **j**, Fit of the TFIIA crystal structure (PDB: 1YTF)[23] to OC2-focused cryo-EM density in a Pol II top view[50] (left). The four-helix bundle undergoes a minor rotation towards the β-barrel, and is apparently flexible (compare Extended Data Fig. 5e). Toa1 (middle) and Toa2 (right) subunit structures are shown. A large non-conserved insertion in Toa1 (Δ95–209), lacking in recombinant TFIIA (Methods), may affect the relative positioning of the four-helix bundle to the β-barrel. **k**, Fit of the TFIIF model to OC cryo-EM densities viewed from the top[50]. TFIIF dimerization domain and Tfg1 N-terminal region, arm, and charged helix elements are superimposed on the OC4 cryo-EM density. Tfg2 linker and WH domains are superimposed on OC2 and OC3-focused cryo-EM densities, respectively. Subunit architectures for Tfg1 (middle) and Tfg2 (right) subunits are shown, indicating disordered regions. Secondary structure elements were labelled according to the crystallographic model of the human RAP30–RAP74 heterodimer[75]. **l**, Fit of the TFIIE model to OC cryo-EM densities shown from the front[50] (left). Models for Tfa1 eWH, E-linker and E-ribbon are superimposed onto OC1 and OC3 densities. Models for Tfa2 WH1 domain, Tfa2 WH2 and E-tether were fitted into OC3-focused and OC3 densities. Tfa1 (middle) and Tfa2 (right) subunits are shown, indicating disordered regions. The connectivity of the E-tether helices remains uncertain. **m**, Fit of promoter DNA to OC cryo-EM densities is shown in a side view[50]. A weak density for single-stranded template DNA contacts the Pol II fork loop 1, and is indicated by a blue arrow. Upstream and downstream DNA models are superimposed with OC3 and OC1 densities, respectively. The location of the Pol II active site magnesium ion is indicated.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 2 | Details of TFIIF and the upstream DNA assembly. a, View of the open complex from the side[50]. Pol II elements external 2 (dark green), lobe (yellow), protrusion (orange), Pol II subunit Rpb12 (dark blue) and basal factors TBP, TFIIB, and TFIIF are coloured as in Fig. 1. The remainder of the open complex is transparent. Green and purple boxes indicate the locations of TFIIB C-terminal cyclin and TFIIF dimerization domains, respectively. **b**, Interactions of TFIIB C-terminal cyclin domain with Pol II protrusion, Rpb12, Tfg2 linker and DNA. Colours as in **a**. **c**, Details of TFIIF dimerization domain interactions with Pol II external 2 and lobe[50]. **d**, Crystallographic analysis of the yeast-specific Tfg1 N-terminal region. Weak density for the Tfg1 N-terminal region was observed by cryo-EM (OC4 round 2 class 2) at low contour level (0.0155) close to Pol II elements external 1 and the hybrid binding region[50] (left). X-ray analysis (right) of the corresponding peptide (Tfg1 F21–R35) enabled modelling and assignment of residue M27 (indicated with asterisk) owing to the anomalous signal. The $F_o - F_c$ electron density map (grey, contour level 2.5σ), seleno-methionine anomalous difference Fourier (yellow, contour level 5σ), and final model in ribbon presentation (purple) are shown. The sequence of the synthetic peptide used for soaking into Pol II crystals is shown below. The modified methionine residue and predicted secondary structure are indicated. **e**, The $F_o - F_c$ electron density maps obtained from soaking Pol II crystals with TFIIF (purple) and seleno-methionine labelled peptide (grey), respectively, show similar density in the same location on Pol II. **f**, The putative Tfg2 C terminus contacts TBP. Viewed from the side[50]. A tubular cryo-EM density from the OC3 map, low-pass filtered to 8 Å, emanates from the TFIIF Tfg2 WH–TFIIE Tfa2 WH1 density, and was tentatively assigned to the Tfg2 C-terminal region. The putative Tfg2 density reaches the TBP subunit, consistent with their suggested interaction[29,76].

© 2016 Macmillan Publishers Limited. All rights reserved

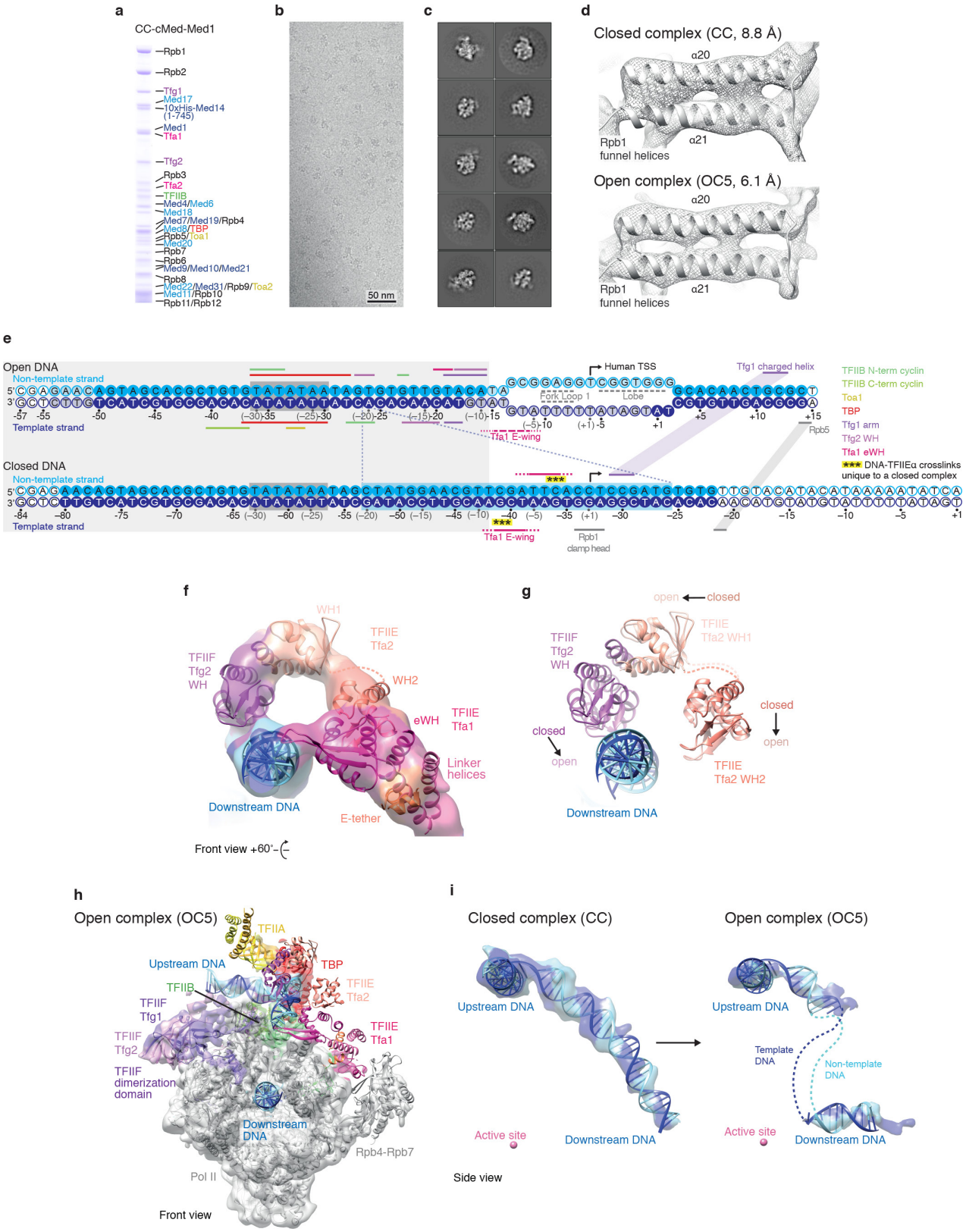


Extended Data Figure 3 | See next page for caption.

© 2016 Macmillan Publishers Limited. All rights reserved

Extended Data Figure 3 | Structure–function analysis of TFIIE and its interactions in the open complex. a, The architectural model of TFIIE contains all regions required for viability in yeast[16]. A domain schematic (top) indicates the good overlap between modelled (dashed line) and functionally essential regions. Essential (grey), partially redundant Tfa2 WH1 and WH2 domains (blue), and non-essential elements (cyan) are indicated on the TFIIE model, shown in previously defined front and top views[50] of Pol II. **b**, TFIIE sequence conservation. The sequence conservation among *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster*, *Gallus gallus*, and *Homo sapiens* was mapped onto a ribbon representation of the TFIIE model. Highly, strongly, weakly and non-conserved residues are coloured in green, yellow, white, and grey, respectively. The location of a non-modelled helical density in the OC3 cryo-EM map, which may correspond to Tfa1 helix α7, is indicated. Views as in **a**. **c**, An additional density (green) in the OC3 cryo-EM map on top of the Tfa1 E-wing was tentatively assigned to Tfa1 helix α7 and this may stabilize the long β-hairpin. A front view is shown[50]. **d**, Tfa1–FeBABE cleavage sites in TFIIE[16] are consistent with the TFIIE architecture. **e**, Tfa1– and Tfa2–FeBABE cleavage sites in the Pol II clamp[16] and a protein–protein crosslink between Rpb1 K212 (Pol II clamp)–Tfa2 K277 (TFIIE E-tether)[17] are consistent with the location of eWH and E-tether. **f**, Tfa2–Tfg2 protein–protein crosslinks[17] are consistent with the Tfg2 WH–Tfa2 WH1 architecture. **g**, The TFIIE mutations used for functional characterization were mapped onto a domain schematic and the model of TFIIE, shown in a front view[50]. **h**, Pulldown assays with recombinant TFIIE variants carrying mutations at the TFIIE–CC interface revealed that the E-ribbon is essential for TFIIE recruitment. For details of the TFIIE mutants, see **g**. Pulldowns were analysed by SDS–PAGE (Coomassie staining). To confirm the integrity of the purified TFIIE variants, 2 μg were analysed (left). Some minor contaminant and degradation bands of TFIIE are indicated by an asterisk. The bead elution from the pulldown assay is shown (middle), providing negative (no TFIIE) and positive (TFIIE) controls in the two leftmost lanes. The binding of all TFIIE variants to the CC was impaired compared to the wild-type protein, with the exception of the Tfa1(ΔE-wing) mutant, suggesting that all other interfaces contribute to TFIIE binding affinity. The most severe binding defect was observed upon mutation of three residues in the E-ribbon (Tfa1(L134/V137/L140)) to glutamate or alanine. This suggests that the E-ribbon is largely responsible for recruitment of TFIIE to the CC. The bead-only control (right) indicated that TFIIE and TFIIE variants did not show unspecific binding to the beads. **i**, Western blot analysis of the 3×Flag-tagged Tfa1 confirms specific immune-depletion of Tfa1 in the nuclear extract (NE), whereas levels of Pol II (Rpb3), TFIIB, and Histone H3 were unaffected. **j**, Yeast complementation assays were performed in triplicate experiments with wild-type *TFA1*, an empty vector, and *TFA1* variants with mutations in the TFIIE eWH domain (N50E/K51E/T52E, N50A/K51A/T52A, and P56A/A59E/R62E in eWH helix α3, and ΔE-wing), or the E-ribbon(L134E/V137E/L140E) (see Methods). **k**, The long E-wing in the TFIIE subunit Tfa1 eWH is characteristic of WH domains involved in DNA strand separation[77]. The upstream edge of the transcription bubble and eWH domain are shown in a front view[50] rotated by ∼20° in the horizontal axis. Corresponding regions of human (*Hs*) Werner syndrome ATP-dependent helicase (WRN) WH (PDB: 2WWY) and RecQ1 WH (PDB: 3AAF) domains are shown.

© 2016 Macmillan Publishers Limited. All rights reserved

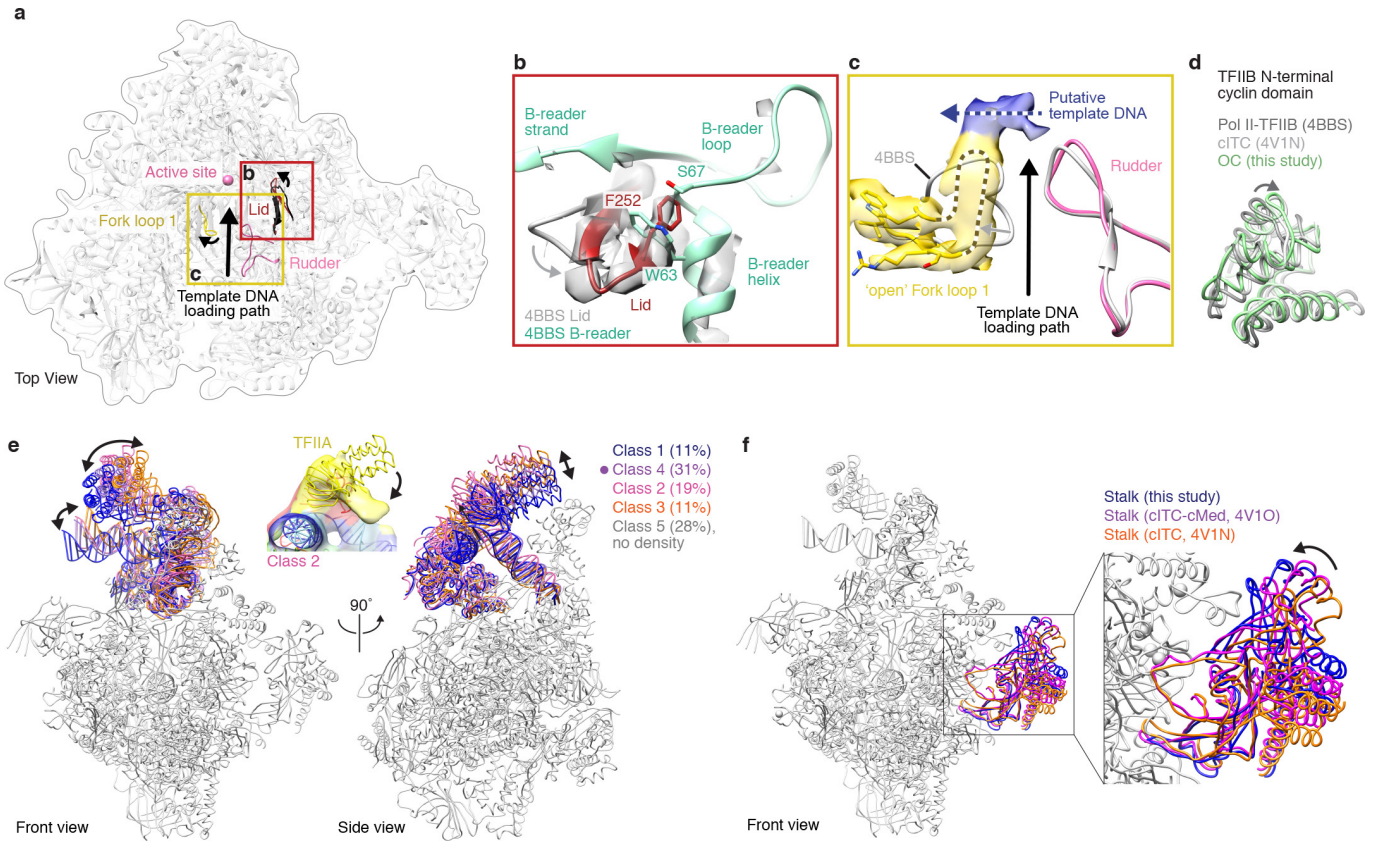


Extended Data Figure 4 | See next page for caption.

© 2016 Macmillan Publishers Limited. All rights reserved

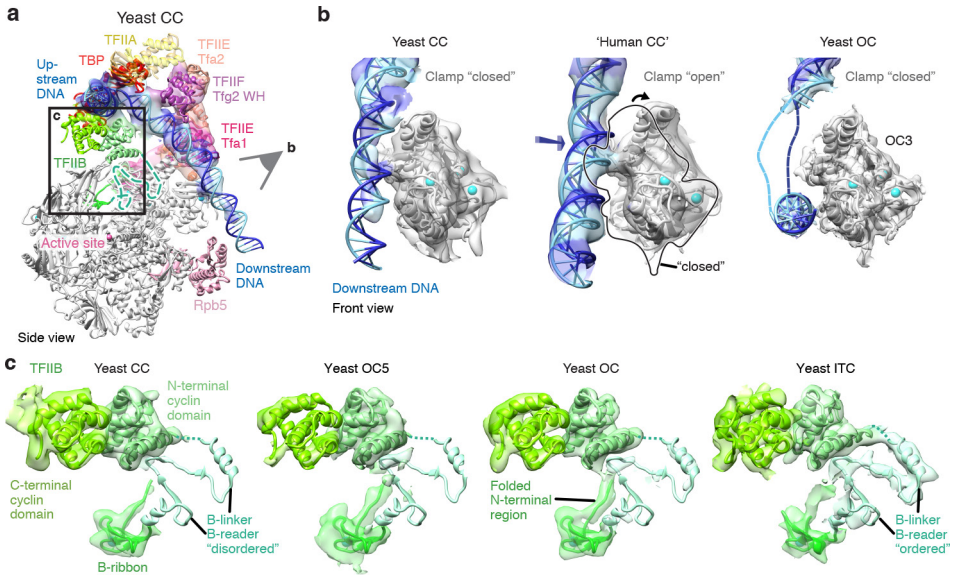
Extended Data Figure 4 | Closed complex and spontaneously formed open complex. a, SDS–PAGE analysis of CC–cMed–Med1 complex after size-exclusion chromatography. Protein colours as in Fig. 1. **b**, Cryo-EM micrograph of the CC–cMed–Med1 complex. Scale bar, 50 nm. **c**, Ten representative reference-free 2D class averages of CC–cMed–Med1 reveal flexibility for the upstream complex. Core Mediator was not retained during cryo-EM analysis. **d**. Detailed view of the Pol II funnel helices in the CC (top) and OC5 (bottom) densities. **e**, Promoter sequences and differences in protein–DNA interactions are shown for the two distinct nucleic acid scaffolds used for preparation of closed and open complexes (compare Fig. 1d). Coloured bars indicate DNA–protein interaction. Solid, shaded, and empty circles respectively represent nucleotides included in the structure, excluded owing to weak cryo-EM density, or excluded owing to absence of cryo-EM density. Analogous yeast (black) and human (grey) numbering of promoter DNA is shown. The TATA-box sequence (red box) and *HIS4*-promoter sequence absent in the modified OC nucleic acid scaffold[19] (grey box) are indicated. Protein–DNA interactions in the region covered by the light grey box are unchanged between CC and OC, and shown only for the OC for clarity. Unique and altered interactions are shown for each complex. DNA–TFIIEα photo-crosslinks, indicated by black asterisks, were observed in a closed but not open promoter state[40] and are consistent with the CC model. **f**, Fit of TFIIE, Tfg2 WH and downstream DNA into CC density. Two rigid bodies were used for fitting: (i) Tfg2 WH and Tfa2 WH1 and (ii) Tfa2 WH2, eWH, E-linker and E-tether helices. Although the overall fit reflects density well, the eWH domain and its E-wing may be rotated further away from promoter DNA. **g**, Details on the location of downstream DNA (template, blue; non-template, light blue), Tfg2 WH, and Tfa2 WH1 and WH2 in the closed (dark colours) and open (light colours) complexes in the same view as in **f**. **h**, Cryo-EM density of OC5 and the OC ribbon model are shown in a front view[50]. The OC5 map shows weak density in regions of upstream assembly, TFIIE, and DNA that may be caused by increased flexibility owing to the heterogeneous population of spontaneously opened DNA. Colours as in Fig. 1. **i**, Fit of promoter DNA to cryo-EM densities of CC and OC5, shown in a side view[50].

© 2016 Macmillan Publishers Limited. All rights reserved



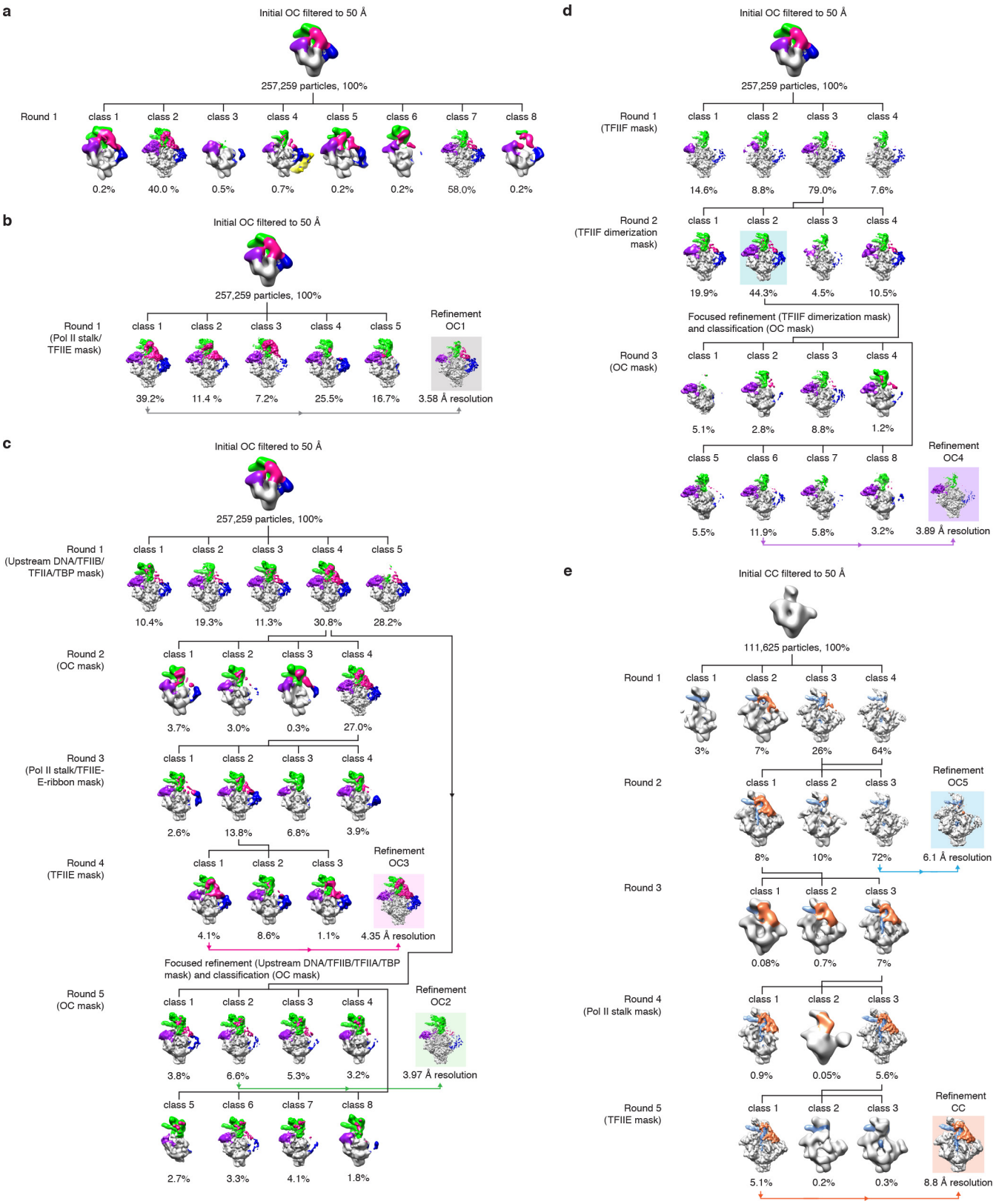
Extended Data Figure 5 | Pol II cleft clearance, structural flexibility and rearrangements in the OC. a, Pol II lid and fork loop 1 assume new conformations in the OC, clearing the Pol II cleft for loading of single-stranded template DNA. Arrows indicate the direction of movement of the two Pol II elements, and the template DNA loading path. The lid (dark red) in the open complex is moved in comparison to the lid of a Pol II–TFIIB ITC crystallographic study (PDB: 4BBS). Yellow and red boxes indicate zoomed-in regions of **b** and **c**, respectively. **b**, The movement in the Pol II lid leads to a steric clash with the TFIIB B-reader, observed in a Pol II–TFIIB ITC crystal (PDB: 4BBS), and facilitates its withdrawal in the open complex. In particular the lid residue F252 clashes with W63 and S67 of the B-reader. The OC1 cryo-EM density is shown for both lid and B-reader elements. **c**, The cryo-EM density of the OC1 reveals an 'open' Pol II fork loop 1 and a stably associated fragment of putative template DNA. The 'open' state of fork loop 1 provides additional space for loading of single-stranded template DNA past the Pol II rudder, towards the active site cleft. **d**, The position of the TFIIB N-terminal cyclin domain (light green) is altered in comparison to a Pol II–TFIIB ITC crystal structure[15] (dark grey), but similar to its location in a cITC[19] (light grey), probably owing to the presence of DNA. **e**, Flexibility of the upstream DNA assembly. The cryo-EM data of the OC was sorted on the basis of structural differences using an upstream assembly mask that included upstream DNA, TFIIA, TBP, and TFIIB cyclin domains (OC2 round 1, compare Extended Data Fig. 7c). Four of five resultant classes revealed different positions of the upstream complex, indicated here by fitted ribbon models of the OC. Previously defined front and side views[50] are shown. Class 2 (middle) revealed the TFIIA four-helix bundle rotated by 85°, consistent with a high degree of flexibility. Class 4 represents the largest fraction of the data (31%), and gave a more defined density for the upstream complex, which was improved by further classification (Extended Data Fig. 7c). Class 5 presented with no density for the upstream complex or the Tfg2 linker, but did show density for the TFIIB B-ribbon and the TFIIF dimerization domain, suggesting that TFIIB and TFIIF remained bound to the complex. This is consistent with TFIIF-dependent association of the TFIIB-core domain with the Pol II wall[27], and this apparently requires an ordered Tfg2 linker. **f**, The Rpb4–Rpb7 stalk adopts different positions in cITC, cITC-cMed, and OC. This suggests that Mediator and TFIIE may bind co-operatively. This is consistent with previous findings[78] and with pulldowns (Extended Data Fig. 3h), which suggest that the TFIIE E-ribbon–stalk interface, which is important for TFIIE recruitment, is stabilized in the presence of Mediator.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 6 | Pol II clamp positions and TFIIB B-reader mobility during DNA opening. a, The yeast CC is shown from a side view[50], indicating the path of DNA and location of TFIIE. The eye symbol (grey) indicates the point of view in **b**. **b**, The Pol II clamp may undergo transitions during DNA opening as indicated. The OC model of the Pol II clamp is shown superimposed on yeast CC (this study), and yeast OC (this study). The OC model Pol II clamp was rigid-body fitted to the human CC cryo-EM density[20] (EMD-2306) and is superimposed. The view is from the front[50]. **c**, The TFIIB B-reader element shows strong density only in the ITC state, suggesting that its mobility in earlier states may be important for maintaining a cleared path for template DNA loading into the Pol II cleft. Ordering of the B-reader may further lead to stabilization of the upstream promoter assembly that is flexible in the OC (Extended Data Figs 5e, 7c). Cryo-EM densities for yeast CC (this work), OC5 (this work), OC (this work), and ITC (EMD-2785) complexes are superimposed on the TFIIB model (PDB: 4BBS for the B-linker and B-reader). As secondary structure elements could not be resolved in the human CC[20], we excluded this cryo-EM density from comparison.

© 2016 Macmillan Publishers Limited. All rights reserved

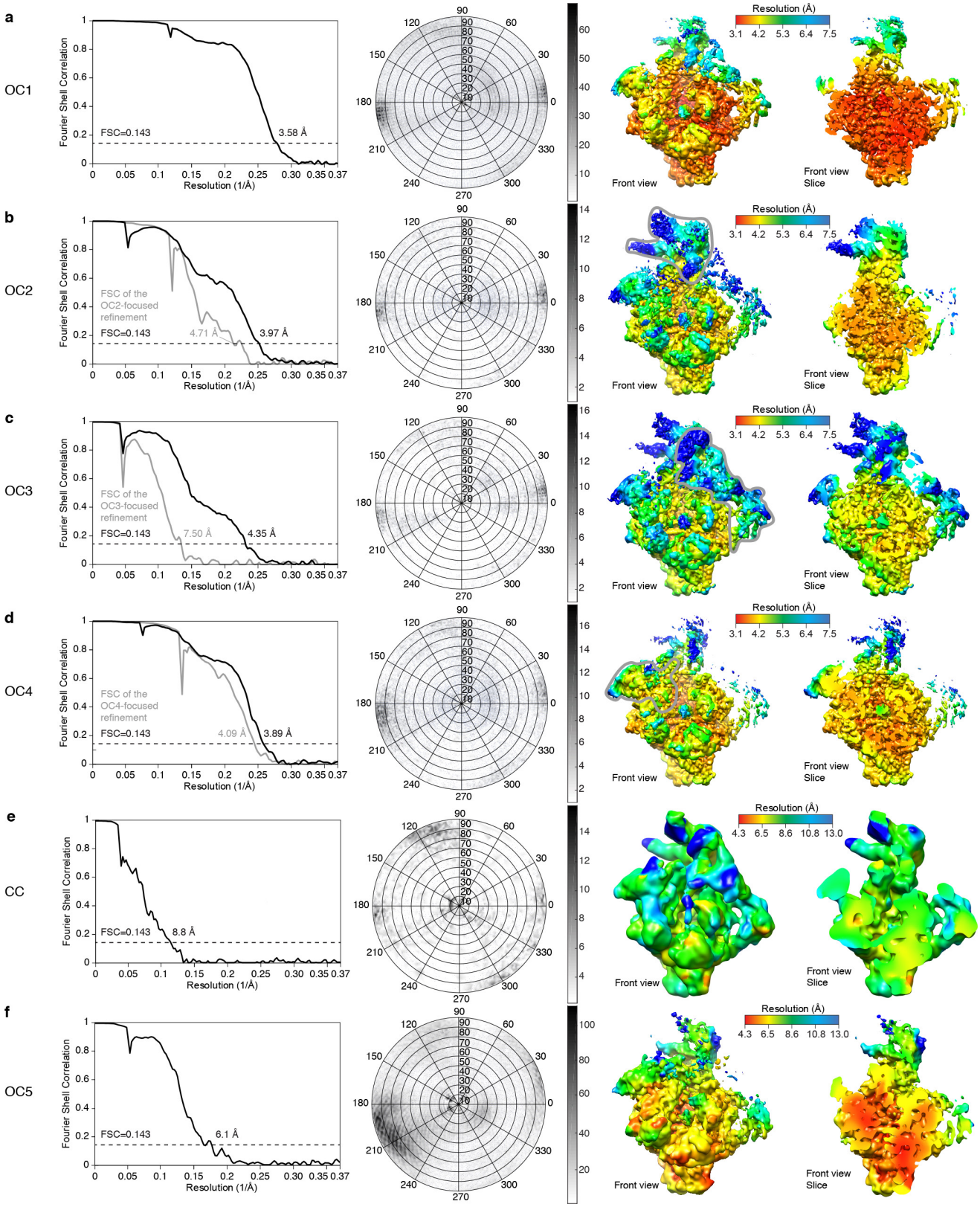


Extended Data Figure 7 | See next page for caption.

© 2016 Macmillan Publishers Limited. All rights reserved

Extended Data Figure 7 | Three-dimensional classification of cryo-EM data. a, Three-dimensional image classification of the cryo-EM data set into eight classes using an initial OC reconstruction as the reference model, revealed heterogeneity. The percentage of single particles contributing to each class is provided. To help visualize structural differences, 3D reconstructions of the OC are coloured according to mobile regions: Pol II core, TFIIB B-ribbon (grey); upstream DNA, TFIIA, TBP, TFIIB cyclin domains, Tfg2 linker (green); TFIIF dimerization domain (purple); TFIIE except E-ribbon, Tfg2 WH (magenta); Pol II Rpb4–Rpb7 stalk and E-ribbon (blue); cMed–Med1 (yellow). **b**, Focused classification into five classes using a mask covering the Pol II stalk and E-ribbon. The resultant class 1 (OC1) was subsequently refined to 3.58 Å resolution (grey box) and revealed the location of the TFIIE E-ribbon. Colours as in **a**. **c**, Improvement of densities for Tfg2 linker, TFIIB, and TFIIE, through rounds of focused 3D classification using various masks. First, heterogeneity due to flexibility of upstream DNA and associated factors was overcome by applying a mask around this region (round 1). Focused refinement of the upstream DNA assembly of the resultant class 4 of round 1 (OC2-focused), improved the density quality for TFIIA (Extended Data Fig. 1j). Classification of the OC2-focused density revealed the upstream DNA complex (OC2) at 3.97 Å resolution (green box). Separate classification of class 4 of round 1 using OC, Pol II stalk and TFIIE E-ribbon, and TFIIE masks yielded class 1 of round 4 (OC3, magenta box) that contained a complete TFIIE density at a nominal resolution of 4.35 Å after 3D refinement (see Extended Data Fig. 8c). The small fraction of stably bound TFIIE is consistent with its reduced affinity to the pre-initiation complex[79]. Focused refinement of OC3 with a TFIIE–stalk mask (OC3-focused) improved density for Tfg2 WH and Tfa2 WH1 domains. Colours as in **a**. **d**, To improve the density of TFIIF dimerization domain and the Tfg1 arm, three rounds of classification using a TFIIF, TFIIF dimerization domain, and OC mask were employed. Class 2 of round 2 (cyan box) enabled fitting of the Tfg1 N-terminal peptide, which was resolved by X-ray analysis (Extended Data Fig. 2d, e). This class was further refined locally using a mask covering the TFIIF dimerization domain, and then classified with an OC mask, revealing class 6 of round 3 at 3.89 Å resolution after 3D refinement (purple box). Colours as in **a**. **e**, 3D classification of the CC cryo-EM data set into four classes, using an initial CC reconstruction as the reference model, revealed heterogeneity. Mobile regions in the reconstructions are highlighted: promoter DNA (blue), TFIIE (except E-ribbon), and Tfg2 WH (orange). Classifying the most populated classes from round 1 into three classes unexpectedly revealed open and closed promoter DNA states in the data set: CC (round 2, class 1) and OC5 (round 2, class 3). Class 3 of round 2 (OC5) was refined to 6.1 Å resolution (blue box). Class 1 from round 2 was further classified into three classes. The resultant class 3 of round 3 revealed density for closed downstream promoter DNA above the Pol II cleft, and TFIIE. The cryo-EM density for downstream DNA and TFIIE was improved by focused classification using two soft-edged masks. A mask covering the Pol II Rpb4–Rpb7 stalk yielded a class with better occupancy for the stalk (round 4, class 3), which was further sorted using a mask covering TFIIE and Tfg2 WH to improve their densities. Class 1 of round 5 was refined to 8.8 Å resolution (CC, orange box).

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 8 | Resolution of cryo-EM reconstructions. **a**, Gold-standard FSC (left) of the OC1 cryo-EM single particle reconstruction (FSC = 0.143). Orientation distribution plot of all particles that contribute to the OC1 reconstruction (middle). The OC1 cryo-EM map is shown (right) from a front view[50] and a central slice through the reconstruction, which are coloured by local resolution as described[19]. **b**, As in **a** but for the OC2 reconstruction. The gold-standard FSC for the density obtained from focused refinement (OC2-focused) with a soft mask around the upstream DNA assembly is indicated in grey (see Methods). The region masked for focused refinement is indicated with a grey outline on the cyro-EM map coloured by local resolution (right). **c**, As in **b**, but for the OC3 and OC3 focus-refined reconstructions. **d**, As in **a**, but for the OC4 and OC4 focus-refined reconstructions. **e**, As in **a**, but for the CC reconstruction. **f**, As in **a**, but for the OC5 reconstruction.

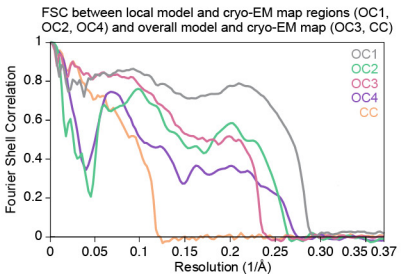
© 2016 Macmillan Publishers Limited. All rights reserved

a **Cryo-EM data collection and refinement statistics of the OC structure**

	Pol II–TFIIB ribbon (OC1)	Upstream assembly* (OC2)	TFIIF dimerization (OC4)
Data collection			
Particles	102,876	17,282	29,455
Pixel Size (Å)	1.35	1.35	1.35
Defocus range (μm)	–0.7 to –4.2	–0.7 to –4.2	–0.7 to –4.2
Voltage (kV)	300	300	300
Electron dose (e^{-} Å^{-2})	33	33	33
Reconstruction (RELION)			
Accuracy of rotations (°)	1.58	1.11	0.96
Accuracy of translations (pixel)	0.71	0.59	0.50
Resolution (Å)	3.58	3.97	3.89
Map sharpening B-factor (Å^2)	111	95	79
Model composition			
Non-hydrogen atoms	27844	3488	1839
Protein residues	3506	444	221
Refinement (PHENIX)			
Map CC (around atoms)	0.768	0.631	0.539
Rms deviations			
Bond lengths (Å)	0.01	0.01	0.01
Bond angles (°)	0.94	1.11	1.25
Validation			
Molprobity score	1.99	2.23	2.51
All-atom clashscore	12.43	18.87	31.30
Rotamer outliers	0.3%	0%	0%
C-beta deviations	0	0	0
Ramachandran plot			
Outliers	0.3%	0.5%	1.0%
Allowed	5.4%	6.8%	8.4%
Favoured	94.3%	92.7%	90.6%

*A model of Upstream assembly components was refined into the OC2 cryo-EM density: Pol II protrusion (residues 44-218 and 405-465), Tfg2 Linker (residues 249-280), TFIIB N-and C-terminal cyclin domains (residues 124-326), and the Tfg1 arm (residues 327-335).

b



c **X-ray crystallographic data collection and refinement statistics**

	Pol II–Tfg1 SeMet	Pol II–TFIIF
Data collection		
Space group	C222(1)	C222(1)
Cell dimensions *a*, *b*, *c* (A)	221.2, 392.7, 282.2	222.8, 392.7, 283.4
Resolution (A)	50-3.5 (3.6-3.5)*	50-3.9 (4.0-3.9)*
R_{sym} (%)	11.3 (105.4)	13.6 (91.3)
$I/\sigma I$	15.7 (2.0)	12.8 (2.4)
Completeness (%)	100 (100)	99.8 (100)
Redundancy	7.1 (6.9)	7.5 (7.6)
$CC_{1/2}$ (%)	99.8 (78.9)	99.8 (82.6)
Refinement		
Resolution (A)	49-3.5	48.9-3.9
No. reflections	151,192	112,717
R_{work}/ R_{free} (%)	15.4/18.0	16.2/19.4
No. atoms		
Protein	31,330	31,330
Ligand/ion	9	9
B-factors (A˚)		
Protein	127.9	143.7
Ligand/ion	125.7	124.4
R.m.s deviations		
Bond lengths (A)	0.01	0.01
Bond angles (°)	1.30	1.26

*Highest resolution shell is shown in parenthesis.

Extended Data Figure 9 | Data collection, refinement statistics, and structure validation. a, Cryo-EM data collection and refinement statistics of the OC structure. Different regions of the composite OC structure were refined into OC1, OC2, and OC4 maps as described (see Methods) to obtain an atomic model for 90% of the structure. **b**, Gold-standard FSC between the respective coordinate models and local regions of the OC1, OC2, and OC4 cryo-EM maps used for model refinement and between overall OC and CC models compared to OC3 (best TFIIE density) and CC cryo-EM maps. **c**, X-ray crystallographic data collection and refinement statistics.

© 2016 Macmillan Publishers Limited. All rights reserved

ARTICLE

doi:10.1038/nature17970

Near-atomic resolution visualization of human transcription promoter opening

Yuan He[1,2], Chunli Yan[3], Jie Fang[4], Carla Inouye[5], Robert Tjian[4,5,6], Ivaylo Ivanov[3] & Eva Nogales[1,4,6]

In eukaryotic transcription initiation, a large multi-subunit pre-initiation complex (PIC) that assembles at the core promoter is required for the opening of the duplex DNA and identification of the start site for transcription by RNA polymerase II. Here we use cryo-electron microscopy (cryo-EM) to determine near-atomic resolution structures of the human PIC in a closed state (engaged with duplex DNA), an open state (engaged with a transcription bubble), and an initially transcribing complex (containing six base pairs of DNA–RNA hybrid). Our studies provide structures for previously uncharacterized components of the PIC, such as TFIIE and TFIIH, and segments of TFIIA, TFIIB and TFIIF. Comparison of the different structures reveals the sequential conformational changes that accompany the transition from each state to the next throughout the transcription initiation process. This analysis illustrates the key role of TFIIB in transcription bubble stabilization and provides strong structural support for a translocase activity of XPB.

Eukaryotic gene transcription is tightly controlled during the initiation stage, when RNA polymerase II (Pol II) and the general transcription factors (GTFs) (TFIID, TFIIA, TFIIB, TFIIE, TFIIF and TFIIH) assemble at the promoter into a PIC[1–3]. Recruitment of TFIID, TFIIA and TFIIB to the DNA facilitates correct loading of the Pol II–TFIIF complex, followed by engagement of TFIIE and TFIIH[4].

The initial closed promoter complex (CC) must transition into an open complex (OC), where the melted single-stranded template DNA is inserted into the active site and Pol II locates the transcription start site (TSS). This transient OC is then converted into an initial transcribing complex (ITC), where the first phosphodiester bond forms and messenger RNA starts to be synthesized[5,6]. Following abortive cycles of synthesis of short RNA products[7], Pol II eventually clears the promoter and a stable elongation complex (EC) forms.

Crystallographic structures and crosslinking data have led to models of partial PIC assemblies for *Saccharomyces cerevisiae*[8–10]. Cryo-EM studies of human[11] and yeast[12,13] PIC complexes have described larger assemblies. Recently, we combined the cryo-EM structures of human TFIID–TFIIA–DNA and a TBP-based PIC to generate a model of the full TFIID-based structure that describes how TFIID recognizes the core promoter and contributes to PIC recruitment[14]. However, the atomic resolution description of this complex machinery and its transitions between the CC, OC, ITC and EC states are yet to be determined.

Here we describe cryo-EM structures at 3.9–8.6 Å resolution leading to near-atomic models of the human TBP-based PIC throughout transcription initiation. Comparisons between states provide new mechanistic insights into the processes of DNA engagement, promoter melting and transcription bubble stabilization.

Visualization of human initiation complexes

We used our previously developed system for assembly and purification of DNA-bound human pre-initiation complexes, in which we used TBP in place of TFIID[11]. We included TFIIS because of its potential role in PIC formation, in addition to its role in elongation[15]. Given that TFIIS can stimulate the weak RNA nuclease activity of Pol II[16,17], we also analysed the ITC structure without TFIIS (denoted hereafter as ITC(−IIS)). In order to mimic the CC, OC and ITC states, we used three different nucleic acid scaffolds (Fig. 1a). This strategy allowed us to visualize the changes in the PIC, from its assembly on duplex DNA, to one preceding clearance of the promoter. Our purified PICs were functionally active (Extended Data Fig. 1). The overall architecture of the four states analysed (Fig. 1b) resembles our reconstructions of CC published previously[11,14].

Local resolution estimation and sorting procedures (Extended data Figs 2, 3) showed that the TFIIH attachment to the rest of the PIC has some flexibility, thus limiting the resolution of the full PIC owing to slight misalignment. To overcome this limitation we also obtained reconstructions of the 'PIC core', in the absence of TFIIH (Fig. 1c). The density maps for the full PIC and the core PIC were reconstructed using the same data set for each state, generated by automatically picking particles from images of samples that included all GTFs, but for which some of the complexes were incomplete. A different strategy was used for the core CC (see Extended Data Fig. 2 versus Extended Data Fig. 4), because for the CC sample, TFIIE was often not present in complexes lacking TFIIH (in agreement with a cooperative mode of incorporation for TFIIE and TFIIH into the PIC[11]). Instead, a local focused refinement of the CC core was conducted for the particles used to reconstruct the full CC (Extended data Fig. 2).

The global resolution for the core OC, ITC and ITC(−IIS) states is ~3.9 Å, and for the CC core is 5.4 Å. Peripheral GTFs are more flexible and at a slightly lower resolution, ranging from 4.5 Å to 7.5 Å. However, Pol II and most of TFIIB were determined beyond 4.5 Å resolution, approaching 3 Å in certain regions (Fig. 2, Extended Data Fig. 4 and Supplementary Videos 1, 2). For the rest of the complex, secondary structure was clearly visible throughout, allowing homology modelling and flexible fitting to eventually produce near-atomic models for the PIC at each step (Fig. 1d and Extended Data Fig. 5). The nucleic acids are clearly visualized, and separation between bases is apparent close to the active site (Fig. 2b and Supplementary Video 2). Interestingly, there is no density for RNA in the presence of TFIIS (ITC), whereas the RNA is clearly visible in the absence of TFIIS (ITC(−IIS))

[1]Molecular Biophysics and Integrative Bio-Imaging Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. [2]Department of Molecular Biosciences, Northwestern University, Evanston, Illinois 60208, USA. [3]Department of Chemistry, Center for Diagnostics and Therapeutics, Georgia State University, Atlanta, Georgia 30302, USA. [4]Howard Hughes Medical Institute, University of California, Berkeley, California 94720, USA. [5]Li Ka Shing Center for Biomedical and Health Sciences, University of California, Berkeley, California 94720, USA. [6]Department of Molecular and Cell Biology, University of California, Berkeley, California 94720, USA.

© 2016 Macmillan Publishers Limited. All rights reserved

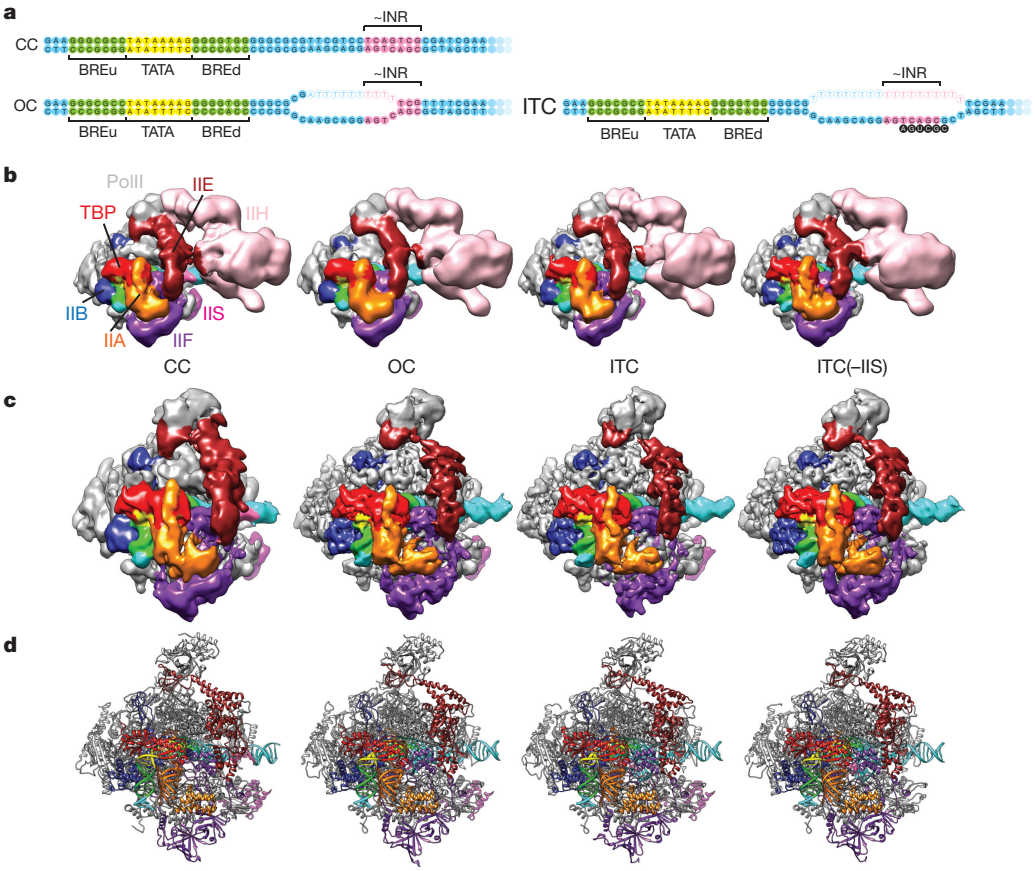


Figure 1 | Cryo-EM reconstructions of human PIC in different states during the initiation process. a, Nucleic acid scaffolds used. Filled and open circles correspond, respectively, to the core promoter and poly-T mismatch sequences. For the OC, two additional bases upstream are opened in the cryo-EM structure. Black circles correspond to RNA. **b**–**d**, Cryo-EM reconstructions of holo-PICs (**b**), core PICs (**c**), and MDFF models (**d**), for the CC, OC, ITC and ITC(−IIS) states.

(Fig. 2b). This result indicates that TFIIS may have changed RNA affinity for the template DNA, caused backtracking of Pol II, or activated Pol II nuclease activity.

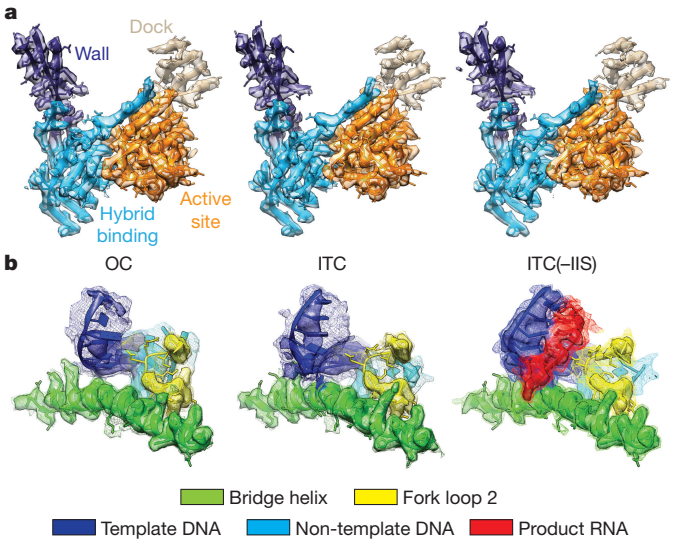


Figure 2 | Examples of near-atomic resolution regions. a, EM densities and corresponding atomic models for the double-ψβ-barrel domain composing the conserved core of Pol II within RPB1 and RPB2 (ref. 49) (see also Supplementary Video 1). **b**, Structural details around the Pol II active site. Densities are shown at two different thresholds. The lower threshold (mesh) allows visualization of the more flexible elements.

Revised structural models for core PIC components

We traced the entire length of TFIIB in the OC, ITC and ITC(−IIS) reconstructions, and most of it in the CC structure (Fig. 3a and Extended Data Fig. 5a). Instead of the TFIIB linker helix described in crystal structures of yeast Pol II–TFIIB[8,18], we observed an extended loop directly contacting the single-stranded non-template DNA. This segment is the least conserved region in TFIIB and may represent a functional departure between yeast and humans. Alternatively, the structure of this region of TFIIB within our PIC complexes could be a result of the interaction with its natural ligand, an initiation DNA bubble, which was absent from the crystallographic studies. In fact, the entire path of the B-linker is slightly different between the human and yeast structures (Fig. 3a).

Our structures revealed an extended interface between the N-terminal lobe of TBP and residues 307–332 in TFIIAβ (Fig. 3b, Extended Data Fig. 5b). This interaction is in agreement with mutagenesis studies[19,20], and was partially captured in a structure of yeast TBP–TFIIA–TATA-box[21]. This conserved surface on TBP is a site for association with transcription activators or repressors[22,23], and has been shown to be critical for releasing the inhibitory effect of TAF1 within TFIID, possibly through binding to TFIIA[24].

The RPB1 trigger loop has been observed in crystallographic structures of the yeast Pol II in either a closed 'on' state, engaged with the last-added nucleotide[25], or an open 'off' state[26]. In all of the initiation states described here, the trigger loop adopts an 'off' state (Fig. 3c, Extended Data Fig. 5c).

Consistent with previous models of TFIIF within the PIC[11,27,28], the dimerization domain of its RAP30 and RAP74 subunits is located near the RPB2 lobe and RPB9 jaw of Pol II, whereas the C-terminal winged-helix (WH) domain in RAP30 directly contacts the BREd

© 2016 Macmillan Publishers Limited. All rights reserved

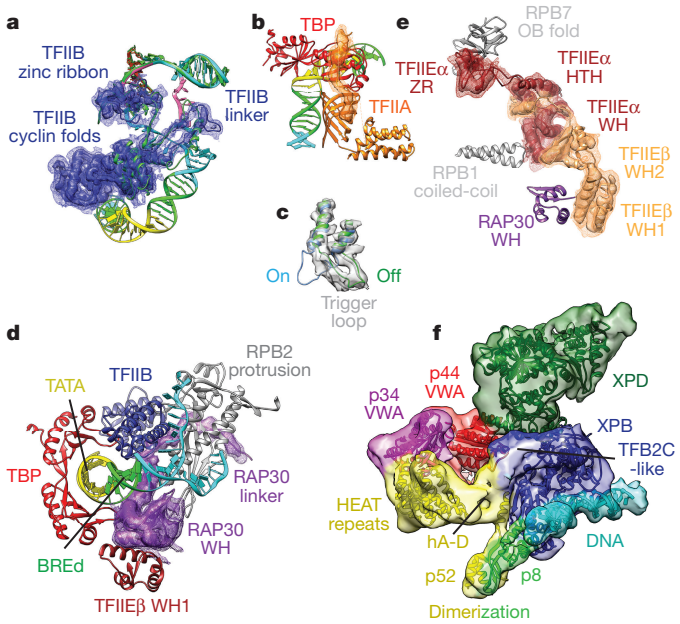


Figure 3 | Newly visualized structural elements of the human PIC. Segmented cryo-EM densities from the ITC(−IIS) reconstruction shown at two different thresholds (the lower threshold (mesh) facilitates visualization of the more flexible elements). **a**, Differences in the path of the B-linker of TFIIB between the yeast crystallographic model[10] (green) and the human ITC cryo-EM structure (coloured as in the rest of the figures). Superimposition was performed by aligning the rigid part of Pol II (that is, excluding the clamp and the stalk), here and for all other figures. **b**, Residues 307–332 of TFIIAβ interact with the N-terminal lobe of TBP. **c**, Trigger loop in RPB1 in an open 'off' state. The trigger loops in crystal structures of yeast EC in its closed 'on' state (2NVZ[25], light blue), and in its open 'off' state (1Y1V[26], green) are shown for comparison. **d**, RAP30 WH domain and linker within TFIIF. Elements of the PIC in close proximity are indicated. **e**, Key domains of TFIIE and their interaction partners within the PIC. **f**, Cryo-EM structure of TFIIH (the flexible CAK subcomplex is not visible after averaging) obtained from a combined data set of open promoter states. Segmentation based on fitting of available structures and homology models.

(downstream TFIIB recognition element) DNA, directly above the RPB2 protrusion (Fig. 3d, and Extended Data Fig. 5e). We modelled the RAP74 α1 helix, an additional α2 helix, and the linker between α1 and the dimerization domain, following a continuous region of density in our maps. Residues 119–175 in RAP30, which constitute the linker between the dimerization and WH domains, were modelled *de novo* for the four states. This linker makes extensive contacts with the RPB2 external 2 and protrusion of Pol II, as well as with TFIIB, TBP, and the BREd, implying that it plays a critical role in the correct positioning of the RAP30 WH domain for stabilization of the promoter DNA. This proposal is consistent with studies in yeast showing that any truncation from the WH end of the linker is lethal for growth[28]. We could not localize the RAP74 arm domain with certainty in any of the states, indicating that this part of TFIIF is probably flexible.

Our density maps allowed us to model the major domains in TFIIE: the WH, zinc ribbon (ZR) and helix-turn-helix (HTH) domains of TFIIEα, and the WH1 and WH2 domains of TFIIEβ (Fig. 3e and Extended Data Fig. 5d). The relative positioning of the three WH domains in TFIIE follows the previous beads-on-a-string model based on XL–MS (protein crosslinking coupled with mass spectroscopy) analysis[9], but the orientation of the domains had to be revised. The first three helices in the WH domain of TFIIEα are involved in dimerization with the WH2 domain of TFIIEβ, whereas the 4th helix directly interacts with the loop at the tip of the RPB1 clamp coiled-coil in the three open promoter states. Further down, the first helix in the WH1 domain of TFIIEβ is involved in the interaction with the WH domain of RAP30. The ZR domain of TFIIEα associates with the deep groove in the RPB7 oligonucleotide/oligosaccharide-binding fold (OB fold) in the stalk. Two helices extending from both sides of the ZR were modelled following helix-like densities directly associated with the WH in TFIIEα. Additional densities near the HTH domain of TFIIEα might correspond to the C-terminal extension of TFIIEβ following the second WH, contributing to an extended dimerization interface between the two TFIIE subunits. This proposal agrees with crosslinking studies indicating that the C terminus of TFIIEβ is close to the region flanking the ZR within TFIIEα, in addition to the WH[9,29]. This C-terminal extension has been shown to be particularly important for viability[9], in agreement with a critical contribution to the dimerization between the two subunits of TFIIE.

Despite previous efforts[30–32], we still lack a detailed subunit architecture for TFIIH. By combining the three open promoter data sets (OC, ITC and ITC(−IIS)) and following a focused refinement procedure (Extended Data Fig. 6), we produced an improved cryo-EM reconstruction of DNA-bound, 'core' TFIIH (not including the mobile kinase module[11]) that allowed the clear visualization of the major and minor grooves of the TFIIH-engaged DNA, and the modelling of six out of the seven core subunits (Fig. 3f). Consistent with our previous model[11], XPD is near the Pol II stalk and TFIIE, whereas XPB is located at the other end of the horseshoe-shaped TFIIH density, engaging the downstream DNA. When comparing our XPB model and the structure of the *Sulfolobus solfataricus* SWI2/SNF2 ATPase core bound to DNA[33], the N-terminal domain of both enzymes interacts with the duplex DNA in a very similar manner (Extended Data Fig. 6b, c). However, the C-terminal domains are rotated with respect to each other (Extended Data Fig. 6d). The rotation resembles the conformational change proposed for the SsoRad54 ATPase[34] on the basis of a FRET (fluorescence resonance energy transfer) study and the crystallographic structure of the protein in the absence of DNA[35]. In our model, the DNA-bound XPB is in a completely different conformation from that captured for SsoRad54, emphasizing the extensive conformational plasticity of these enzymes, probably related to the ATP-driven changes that result in DNA translocation.

The crystal structure of P8 and the C-terminal dimerization domain of P52 fits well into our reconstruction, forming a heterotrimer with XPB on one side of the complex (Fig. 3f). We propose that the P52 Hub-A domain (hA-D)[32] interacts with the N-terminal, TFB2C-like domain of XPB (not built into the density), whereas P8 interacts with the XPB C-terminal domain. This model is consistent with the stimulation of XPB ATPase activity by both P52 and P8 (refs 36, 37). The N-terminal, HEAT repeat domain in P52 serves as a bridge between the two ends of TFIIH. It directly interacts with P34, which in turn contacts the P44–XPD subcomplex on the other side of TFIIH. Because of the lower resolution for this region, possibly owing to flexibility, the orientations of the P34 von Willebrand A (VWA) domain and the P52 HEAT repeat are tentative. There is still unexplained density that probably corresponds to P62 and the C-terminal ring finger domains of both P34 and P44. We propose that the P62 PH domain may correspond to a flexible density observed near TFIIE for our four reconstructions of the full PIC (Extended Data Fig. 6e).

Comparison of CC structure with previous yeast models

The crystallographic structures of yeast Pol II–TFIIB[8,18], TFIIB–TBP–TATA DNA[38], and TFIIA–TBP–TATA DNA[21] can be combined to generate a model of the Pol II–TFIIB–TFIIA–TBP and DNA complex. In this 'synthetic' structure, the TBP–TFIIA–TFIIB subcomplex and Pol II clamp and stalk regions need to be rotated to fit the human CC cryo-EM map (Extended Data Fig. 7a–c). These three modules appear inter-connected: movement of TBP–TFIIA–TFIIB–DNA closer to Pol II would cause a clash of the DNA with the clamp, thereby pushing the clamp open; opening of the clamp results in tilting of the stalk and engagement of the DNA duplex. As the axis for the TBP–TFIIA–TFIIB–DNA module rotation is by the N-terminal cyclin fold of TFIIB

© 2016 Macmillan Publishers Limited. All rights reserved

(Extended Data Fig. 7a), we speculate that TFIIB in the crystal structure with Pol II may be in a DNA pre-engagement state. We also observe differences between our structure and a previously published yeast cryo-EM CC model[13] (Extended Data Fig. 7d), which itself differs from the synthetic model based on crystal structures. The TBP–TFIIA–TFIIB sub-complex in the yeast CC cryo-EM study is rotated away from Pol II, which does not contact the DNA. Furthermore, the clamp is closed down, as we see for our OC and ITC complexes (see below) or that observed for EC structures[26,39]. These discrepancies may reflect known differences between the yeast and human systems. Whereas the TSS is found ~30 base pairs (bp) from the TATA box in metazoans, this distance can be as long as 120 bp in yeast[40,41]. We speculate that the yeast arrangement would allow more single-stranded DNA to be accommodated by the PIC during scanning for the TSS.

Remodelling of the PIC during promoter opening

The most notable change between the core CC and OC structures, in addition to the DNA itself (see below), is the movement of the Pol II clamp and of TFIIE, which directly contacts the clamp (Fig. 4a). The TBP–TFIIA–TFIIB–TFIIF sub-complex, defining a structural unit that surrounds the upstream core promoter elements, moves slightly, maybe as a relaxation process following the release of tension after the downstream bent DNA is melted (Fig. 4a). The bridge helix does not change during bubble opening (Fig. 2b, Fig. 4b). However, the non-template DNA near the downstream bubble junction clearly tilted the Pol II fork loop 2 in the OC, which might therefore function as a sensor of the open promoter (Fig. 2b, Fig. 4b, c). This loop stays tilted in both the ITC and ITC(−IIS) states, further supporting our hypothesis (Fig. 2b). A very similar conformation has been recently reported for the bacterial ITC[42] (Extended Data Fig. 8), suggesting the functional conservation of this critical element at the initiation stage of transcription across different kingdoms.

The most significant changes during promoter opening concern the Pol II clamp and TFIIE (Fig. 4d and Extended Data Fig. 9b). In the CC, the clamp coiled-coil domain associates with the C-terminal extension of TFIIEβ. However, in the OC, as the coiled-coil moves down when the clamp closes, the connection with TFIIEβ is broken and another interface is established with the 4th helix of TFIIEα, shifting the entire WH by ~5 Å. This change probably makes the interaction with TFIIE more energetically favourable, explaining the increase in affinity of TFIIE observed in the OC and ITC compared to the CC.

In the CC, Pol II interacts directly with the promoter DNA duplex via the 3-strand β-sheet of the clamp head (Fig. 4e, Extended Data Fig. 9c) and the 2-helix bundle in RPB5, (Fig. 4f, Extended Data Fig. 9d), immediately upstream and downstream of the initiator (INR) element (and thus the TSS). In the OC, the clamp head and RPB5 still engage the DNA, but in a slightly different manner (Fig. 4e, f and Extended Data Figs 9c, d).

For the elements described, there are no significant changes between the OC and the ITC and ITC(−IIS) states (Extended Data Figs 9a–d). Notably, the CC reconstruction does not show any detectable density for the TFIIB linker, indicating that it is disordered before promoter opening (Extended Data Figs 5a, 9a). However, in the OC, ITC and ITC(−IIS) structures, the B-linker becomes stabilized by direct interaction with the clamp coiled-coil and rudder on one side, and the non-template DNA strand on the other (Fig. 4c, e and Extended Data Figs 5a, 9a). This intricate protein–DNA network likely plays a key role in stabilizing this part of the initiation bubble, which is fairly solvent accessible.

DNA transition through initiation into elongation

The full promoter DNA is clearly visible in all our reconstructions, making it possible to compare its path throughout these transcription stages (Fig. 5). The upstream DNA, which includes BREu, TATA, and BREd elements, is tightly engaged during all stages of transcription initiation by the TBP–TFIIA–TFIIB–TFIIF sub-complex (Fig. 1c, d).

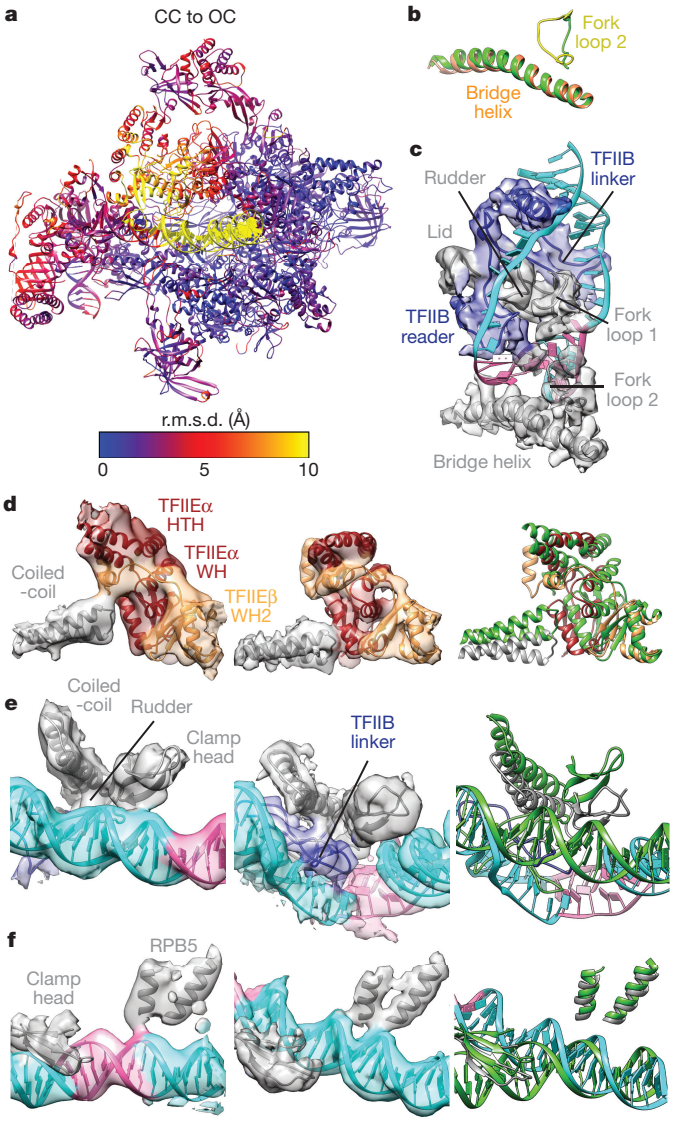


Figure 4 | Structural transitions during promoter opening. a, Colour-coded motion of backbone atoms between CC and OC based on r.m.s.d. **b**, Change of fork loop 2 near the active site between the CC (green) and OC (coloured) (see also Fig. 2b). A similar colour scheme is used for the superposition of CC and OC states in the third column of **d**–**f**. **c**, Segmented cryo-EM density from the OC reconstruction and the corresponding model for TFIIB and various Pol II loops critical for stabilization of the transcription bubble. The density for the nucleic acids has been omitted for clarity. **d**, Changes in the contact between the RPB1 coiled-coil and TFIIE during bubble opening. **e**, Changes near the melting start site involving the clamp head, coiled-coil domain, rudder domain and TFIIB. **f**, Changes in Pol II interaction with downstream DNA involving the clamp head and RPB5.

After the BREd, the path of the promoter in the CC deviates from B-form DNA due to contact with the clamp and RPB5 around the TSS, and further downstream by XPB within TFIIH (Fig. 5a). The effect of XPB binding can be inferred by comparing the present CC structure with our previous cryo-EM structure of core CC in the absence of TFIIH[11] (Fig. 5a). TFIIH causes both a kink at the RPB5 contact, and a slight bend of the DNA at the site of contact with XPB (Fig. 5a, b; see also Fig. 3f). The PIC conformation maintaining this TFIIH-engaged DNA state is probably under strain, primed for DNA unwinding by the ATPase activity of XPB (see below).

In the OC cryo-EM map, two continuous densities connecting the upstream and downstream duplex DNA correspond to the separated

© 2016 Macmillan Publishers Limited. All rights reserved

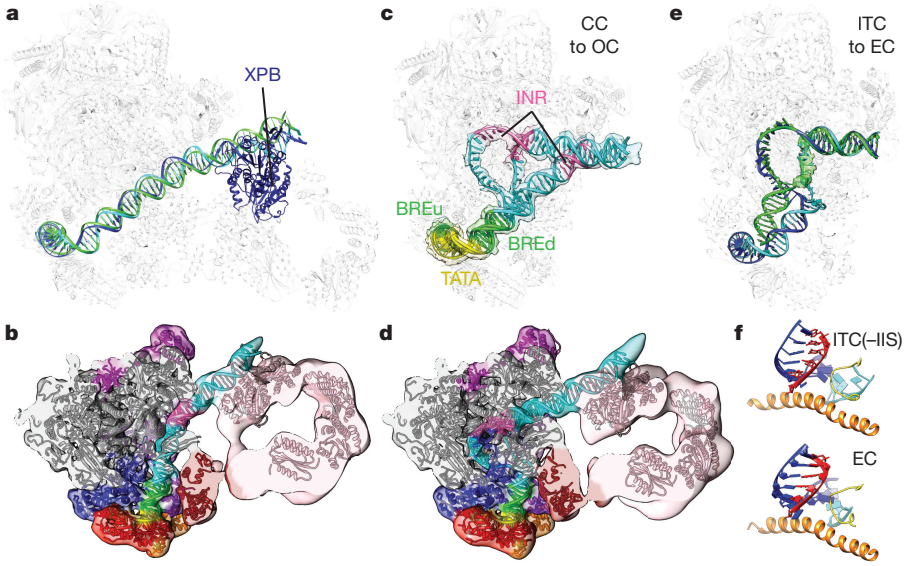


Figure 5 | Nucleic acids rearrangements between stages of transcription. **a**, DNA paths before[11] (green) and after (blue/cyan) TFIIH incorporation into the CC. XPB is coloured navy blue and the rest of the PIC is shown in transparency. **b**, Section through the CC density map and model showing the path of the DNA and its engagement by TFIIH. **c**, Segmented electron microscopy densities and models of promoter DNA in the CC and OC. **d**, Section through the OC density map and model showing the path of the DNA and its engagement by TFIIH. **e**, Comparison of the DNA path in our ITC(−IIS) state (blue/cyan) with that in the bovine EC structure[45] (green). Segmented electron microscopy density of the EC single-stranded non-template DNA is shown to illustrate its path, as this part was not modelled in the published structure. **f**, Close-up views at the active site comparing the position of fork loop 2 relative to the nucleic acids and bridge helix in the human ITC(−IIS) (top) and bovine EC (bottom).

DNA strands (Fig. 5c). The template strand follows the previously characterized tunnel[18] defined by TFIIB, and the wall, rudder, and fork loops of Pol II, whereas the non-template DNA is mainly stabilized by the B-linker and the Pol II fork loops (Fig. 4c and Extended Data Fig. 9a, c). Interestingly, while we used an 11 bp mismatch DNA scaffold for the OC, from position −9 to +2 relative to the TSS (Fig. 1a), incorporation of the DNA into the complex caused a further opening of the bubble by two bps, to position −11 (Extended Data Fig. 9e).

The most notable change between the CC and OC structures concerns the DNA itself (Fig. 4a), beginning ~20 bp downstream of the TATA box, where melting is known to start[43] (Fig. 5c). The upstream DNA barely moves. The location of the 90° bend of the DNA by TBP, further stabilized by TFIIA, TFIIB and TFIIF, likely reinforces and maintains the relative position of the immobilized upstream DNA with respect to Pol II, while a force is applied by TFIIH at the downstream end of the promoter. The minor groove of the DNA contacted by RPB5 superimposes nearly perfectly for the CC and OC reconstructions, whereas the major grooves flanking this point change to a different extent (Figs 4f, 5c). In the CC, there is a bend of the DNA at the RPB5 contact point, whereas the DNA upstream from that point of contact more closely resembles a B-form configuration than previously captured in the absence of TFIIH[11] (Fig. 5a). In the OC, the double-stranded DNA between the active site and the contact with XPB has turned as it slides past the RPB5 contact (Fig. 5c, d). In the process, the kink of the DNA by RPB5 disappears and with it probably also the tension that such a state may impose on the PIC. In contrast, the DNA downstream of the RPB5 contact does not change greatly between the CC and OC states, and XPB maintains an unchanged engagement with DNA.

There are no detectable changes in the double-stranded DNA path between the OC and the ITC, either upstream or downstream of the transcription bubble. The paths of the single-stranded template DNA are more similar than those of the non-template DNA (Fig. 6). Furthermore, whereas the template DNA in both states is visible as a continuous density, there are obvious breaks in the density corresponding to the non-template strand for the ITC. We propose that such breaks in density might be caused by disorder where scrunching is occurring.

Through all the states analysed, the downstream DNA engaged by XPB is in a duplex conformation (Fig. 3f), in agreement with previous translocase models for XPB[9,11,44]. The superimposition of the CC and OC states shows that XPB must translocate by 12 bp along the DNA to position the TSS into the active site of Pol II (Fig. 5c). The same comparison reveals a net rotation of TFIIH by ~10° with respect to the core PIC between the two states (Extended Data Fig. 3j and Supplementary Video 3). The rotational axis passes through two critical contact points: one near the TFIIE–TFIIH interface and the other by the DNA–RPB5 interaction. TFIIH is flexibly associated to the core PIC in each of the states described (Extended Data Figs 2, 3), enabling XPB to reach the downstream DNA while TFIIH maintains contact with TFIIE. As mentioned above, initial binding of TFIIH already causes a small rearrangement of the DNA (Fig. 5a) that may prepare the promoter for melting

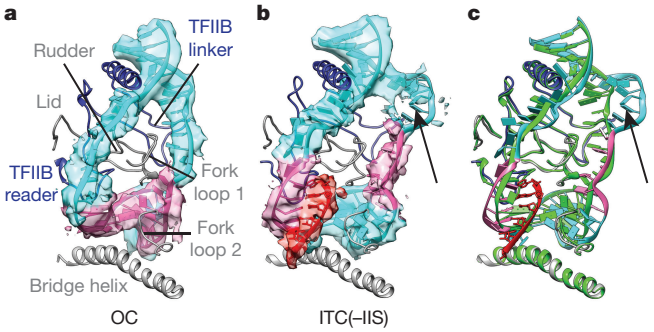


Figure 6 | Possible region of DNA scrunching during initiation. **a**, **b**, Segmented cryo-EM densities and models for DNA and RNA near the Pol II active site for the OC (**a**) and ITC(−IIS) (**b**). TFIIB and Pol II loops that are critical for stabilization of the bubble are also shown (density omitted for clarity). **c**, Superimposition of the OC (green) and ITC(−IIS) (coloured) states. The lack of visible density (arrow) is interpreted as disorder at the site of scrunching, upstream of the contact with the TFIIB linker.

© 2016 Macmillan Publishers Limited. All rights reserved

at the correct position. Pulling of TFIIH by this strained DNA results in a slightly different trajectory of motion for the CC compared to the other states (Extended Data Figs 2, 3). Once the OC state is reached, the range of motion of TFIIH changes, and its average position rotates as indicated in Extended Data Fig. 3j. Through the transition, TFIIH would push the DNA past the RPB5 contact, which thus serves as a pivot point for both DNA and TFIIH (Supplementary Video 3). Within the context of this flexible attachment of TFIIH, the preservation of the XPB–DNA contact, at a similar position with respect to the rest of the PIC for all the states analysed, is in good agreement with a translocation model in which the opening of the transcription bubble occurs by XPB threading the DNA into the Pol II cleft, past the RPB5 contact (Fig. 5b, d, Supplementary Video 3).

When we compare our ITC(−IIS) with a recent cryo-EM structure of a bovine EC[45], the RNA–DNA hybrid and the downstream DNA superimpose, but there are significant differences for the rest of the DNA (Fig. 5e). The upstream duplex DNA in the EC is collapsed by 10 bp (RNA was normalized to the same length), whereas the path is shifted by ~15 Å towards TFIIB, to the location originally occupied by the N-terminal cyclin fold. This shift would happen at the point when the growing RNA has extended to ~12–13 nucleotides[46] and starts to displace TFIIB. Breaking contact with TFIIB, the major connection of Pol II with the rest of the PIC, should thereafter allow Pol II to clear the promoter.

The single-stranded DNA follows slightly different paths for the ITC and EC, partly because of the notable difference in the upstream duplex. Before promoter clearance, extra template DNA has to be fitted within the Pol II cleft, probably resulting in scrunching in the ITC[47,48]. This scrunching could explain the differences we observe for the non-template DNA between our ITC and the EC. The conformation of fork loop 2 in the EC is identical to what we observe in the OC, ITC and ITC(−IIS) states (Figs 5f, 2b, 4b), indicating that this element, probably important for sensing the CC to OC transition by interacting with the single-stranded non-template DNA, remains tilted during the OC to ITC to EC transitions.

Our cryo-EM study, which provides nearly complete pseudo-atomic models of all the structural elements within a functional human PIC, has allowed us to define the structural transitions through the processes of DNA engagement by the CC, opening of the transcription bubble in the OC, and initiation of transcription in the ITC. The present structures constitute a comprehensive structural framework for past and future studies of the complex process of eukaryotic transcription initiation.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 19 January 2015; accepted 5 April 2016.
Published online 11 May 2016.

1. Roeder, R. G. The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem. Sci.* **21,** 327–335 (1996).
2. Goodrich, J. A., Cutler, G. & Tjian, R. Contacts in context: promoter specificity and macromolecular interactions in transcription. *Cell* **84,** 825–830 (1996).
3. Boeger, H. *et al.* Structural basis of eukaryotic gene transcription. *FEBS Lett.* **579,** 899–903 (2005).
4. Buratowski, S., Hahn, S., Guarente, L. & Sharp, P. A. Five intermediate complexes in transcription initiation by RNA polymerase II. *Cell* **56,** 549–561 (1989).
5. Cheung, A. C. M., Sainsbury, S. & Cramer, P. Structural basis of initial RNA polymerase II transcription. *EMBO J.* **30,** 4755–4763 (2011).
6. Liu, X., Bushnell, D. A., Silva, D. A., Huang, X. & Kornberg, R. D. Initiation complex structure and promoter proofreading. *Science* **333,** 633–637 (2011).
7. Luse, D. S. & Jacob, G. A. Abortive initiation by RNA polymerase II *in vitro* at the adenovirus 2 major late promoter. *J. Biol. Chem.* **262,** 14990–14997 (1987).
8. Liu, X., Bushnell, D. A., Wang, D., Calero, G. & Kornberg, R. D. Structure of an RNA polymerase II-TFIIB complex and the transcription initiation mechanism. *Science* **327,** 206–209 (2010).
9. Grünberg, S., Warfield, L. & Hahn, S. Architecture of the RNA polymerase II preinitiation complex and mechanism of ATP-dependent promoter opening. *Nat. Struct. Mol. Biol.* **19,** 788–796 (2012).
10. Sainsbury, S., Niesser, J. & Cramer, P. Structure and function of the initially transcribing RNA polymerase II–TFIIB complex. *Nature* **493,** 437–440 (2013).
11. He, Y., Fang, J., Taatjes, D. J. & Nogales, E. Structural visualization of key steps in human transcription initiation. *Nature* **495,** 481–486 (2013).
12. Plaschka, C. *et al.* Architecture of the RNA polymerase II–Mediator core initiation complex. *Nature* **518,** 376–380 (2015).
13. Murakami, K. *et al.* Structure of an RNA polymerase II preinitiation complex. *Proc. Natl Acad. Sci. USA* **112,** 13543–13548 (2015).
14. Louder, R. K. *et al.* Structure of promoter-bound TFIID and model of human pre-initiation complex assembly. *Nature* **531,** 604–609 (2016).
15. Kim, B. *et al.* The transcription elongation factor TFIIS is a component of RNA polymerase II preinitiation complexes. *Proc. Natl Acad. Sci. USA* **104,** 16068–16073 (2007).
16. Wind, M. & Reines, D. Transcription elongation factor SII. *BioEssays* **22,** 327–336 (2000).
17. Fish, R. N. & Kane, C. M. Promoting elongation with transcript cleavage stimulatory factors. *Biochim. Biophys. Acta* **1577,** 287–307 (2002).
18. Kostrewa, D. *et al.* RNA polymerase II-TFIIB structure and mechanism of transcription initiation. *Nature* **462,** 323–330 (2009).
19. Buratowski, S. & Zhou, H. Transcription factor IID mutants defective for interaction with transcription factor IIA. *Science* **255,** 1130–1132 (1992).
20. Lee, D. K., DeJong, J., Hashimoto, S., Horikoshi, M. & Roeder, R. G. TFIIA induces conformational changes in TFIID via interactions with the basic repeat. *Mol. Cell. Biol.* **12,** 5189–5196 (1992).
21. Bleichenbacher, M., Tan, S. & Richmond, T. J. Novel interactions between the components of human and yeast TFIIA/TBP/DNA complexes. *J. Mol. Biol.* **332,** 783–793 (2003).
22. Juo, Z. S., Kassavetis, G. A., Wang, J., Geiduschek, E. P. & Sigler, P. B. Crystal structure of a transcription factor IIIB core interface ternary complex. *Nature* **422,** 534–539 (2003).
23. Wollmann, P. *et al.* Structure and mechanism of the Swi2/Snf2 remodeller Mot1 in complex with its substrate TBP. *Nature* **475,** 403–407 (2011).
24. Anandapadamanaban, M. *et al.* High-resolution structure of TBP with TAF1 reveals anchoring patterns in transcriptional regulation. *Nat. Struct. Mol. Biol.* **20,** 1008–1014 (2013).
25. Wang, D., Bushnell, D. A., Westover, K. D., Kaplan, C. D. & Kornberg, R. D. Structural basis of transcription: role of the trigger loop in substrate specificity and catalysis. *Cell* **127,** 941–954 (2006).
26. Kettenberger, H., Armache, K. J. & Cramer, P. Complete RNA polymerase II elongation complex structure and its interactions with NTP and TFIIS. *Mol. Cell* **16,** 955–965 (2004).
27. Chen, Z. A. *et al.* Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry. *EMBO J.* **29,** 717–726 (2010).
28. Eichner, J., Chen, H. T., Warfield, L. & Hahn, S. Position of the general transcription factor TFIIF within the RNA polymerase II transcription preinitiation complex. *EMBO J.* **29,** 706–716 (2010).
29. Ohkuma, Y., Hashimoto, S., Wang, C. K., Horikoshi, M. & Roeder, R. G. Analysis of the role of TFIIE in basal transcription and TFIIH-mediated carboxy-terminal domain phosphorylation through structure-function studies of TFIIE-alpha. *Mol. Cell. Biol.* **15,** 4856–4866 (1995).
30. Schultz, P. *et al.* Molecular structure of human TFIIH. *Cell* **102,** 599–607 (2000).
31. Gibbons, B. J. *et al.* Subunit architecture of general transcription factor TFIIH. *Proc. Natl Acad. Sci. USA* **109,** 1949–1954 (2012).
32. Luo, J. *et al.* Architecture of the human and yeast general transcription and DNA repair factor TFIIH. *Mol. Cell* **59,** 794–806 (2015).
33. Dürr, H., Körner, C., Müller, M., Hickmann, V. & Hopfner, K. P. X-ray structures of the Sulfolobus solfataricus SWI2/SNF2 ATPase core and its complex with DNA. *Cell* **121,** 363–373 (2005).
34. Lewis, R., Dürr, H., Hopfner, K. P. & Michaelis, J. Conformational changes of a Swi2/Snf2 ATPase during its mechano-chemical cycle. *Nucleic Acids Res.* **36,** 1881–1890 (2008).
35. Thomä, N. H. *et al.* Structure of the SWI2/SNF2 chromatin-remodeling domain of eukaryotic Rad54. *Nat. Struct. Mol. Biol.* **12,** 350–356 (2005)
36. Coin, F. *et al.* p8/TTD-A as a repair-specific TFIIH subunit. *Mol. Cell* **21,** 215–226 (2006).
37. Coin, F., Oksenych, V. & Egly, J. M. Distinct roles for the XPB/p52 and XPD/p44 subcomplexes of TFIIH in damaged DNA opening during nucleotide excision repair. *Mol. Cell* **26,** 245–256 (2007).
38. Tsai, F. T. F. & Sigler, P. B. Structural basis of preinitiation complex assembly on human pol II promoters. *EMBO J.* **19,** 25–36 (2000).
39. Gnatt, A. L., Cramer, P., Fu, J., Bushnell, D. A. & Kornberg, R. D. Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 Å resolution. *Science* **292,** 1876–1882 (2001).
40. Buratowski, S., Sopta, M., Greenblatt, J. & Sharp, P. A. RNA polymerase II-associated proteins are required for a DNA conformation change in the transcription initiation complex. *Proc. Natl Acad. Sci. USA* **88,** 7509–7513 (1991).
41. Struhl, K. Promoters, activator proteins, and the mechanism of transcriptional initiation in yeast. *Cell* **49,** 295–297 (1987).
42. Zhang, Y. *et al.* Structural basis of transcription initiation. *Science* **338,** 1076–1080 (2012).
43. Giardina, C. & Lis, J. T. DNA melting on yeast RNA polymerase II promoters. *Science* **261,** 759–762 (1993).
44. Kim, T. K., Ebright, R. H. & Reinberg, D. Mechanism of ATP-dependent promoter melting by transcription factor IIH. *Science* **288,** 1418–1421 (2000).

© 2016 Macmillan Publishers Limited. All rights reserved

45. Bernecky, C., Herzog, F., Baumeister, W., Plitzko, J. M. & Cramer, P. Structure of transcribing mammalian RNA polymerase II. *Nature* **529,** 551–554 (2016).
46. Čabart, P., Újvári, A., Pal, M. & Luse, D. S. Transcription factor TFIIF is not required for initiation by RNA polymerase II, but it is essential to stabilize transcription factor TFIIB in early elongation complexes. *Proc. Natl Acad. Sci. USA* **108,** 15786–15791 (2011).
47. Kapanidis, A. N. *et al.* Initial transcription by RNA polymerase proceeds through a DNA-scrunching mechanism. *Science* **314,** 1144–1147 (2006).
48. Revyakin, A., Liu, C., Ebright, R. H. & Strick, T. R. Abortive initiation and productive initiation by RNA polymerase involve DNA scrunching. *Science* **314,** 1139–1143 (2006).
49. Iyer, L. M., Koonin, E. V. & Aravind, L. Evolutionary connection between the catalytic subunits of DNA-dependent RNA polymerases and eukaryotic RNA-dependent RNA polymerases and the origin of RNA polymerases. *BMC Struct. Biol.* **3,** 1 (2003).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank S. Zheng for providing XPB mAb, D. King for providing peptides, P. Grob and T. Houweling for electron microscopy and computer support, respectively; S. Scheres for advice on image processing, S. Kassube for providing TFIIS, and G. Cardone for advice concerning filtering according to local resolution. This work was funded by NIGMS (GM63072 to E.N. and GM110387 to I.I.) and the National Science Foundation (MCB-1149521 to I.I.). Computational resources were provided in part by XSEDE (CHE110042) and the National Energy Research for Scientific Computing Center (DE-AC02-05CH11231). E.N. and R.T are Howard Hughes Medical Institute Investigators.

Author Contributions Y.H. designed and carried out the experiments; C.Y. and I.I. performed structural modelling; C.I., J.F. and Y.H. purified GTFs and Pol II; Y.H. and E.N. analysed the data and wrote the paper.

Author Information Cryo-EM density maps have been deposited in the Electron Microscopy Data Bank (EMDB) under accession numbers EMD-8132 (OC), EMD-8133 (ITC), EMD-8134 (ITC(−IIS)), EMD-8135 (CC core), EMD-8136 (OC core), EMD-8137 (ITC core), EMD-8138 (ITC(−IIS) core), EMD-8131 (TFIIH core). Model coordinates have been deposited in the Protein Data Bank (PDB) under accession numbers 5IY6 (CC), 5IY7 (OC), 5IY8 (ITC), 5IY9 (ITC(−IIS)), 5IYA (CC core), 5IYB (OC core), 5IYC (ITC core), 5IYD (ITC(−IIS) core), 5IVW (TFIIH core). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Y.H. (yuanhe@northwestern.edu) and E.N. (ENogales@lbl.gov).

© 2016 Macmillan Publishers Limited. All rights reserved

METHODS

Data reporting. No statistical methods were used to predetermine sample size.

Transcription complex assembly and purification. TBP, TFIIA, TFIIB, TFIIE, TFIIF and TFIIS were recombinantly expressed and purified from *Escherichia coli*. Pol II and TFIIH were immunopurified from HeLa cell nuclei (that tested negative for mycoplasma) following previously established protocols[50,51]. The design of the DNA construct was based on the super core promoter[52], except that a BREu element was introduced upstream of the TATA box[38] and an EcoRI restriction enzyme site was included downstream of the INR element for purification purposes (template, 5′-ACTGGGGAATTCCATGGTCCGTAGGCACGTCTGCTCGGCT CGAGTGTTCGATCGCGACTGAGGACGAACGCGCCCCCACCCCCTTTTA TAGGCGCCCTTC-3′; non-template 1, 5′- GAAGGGCGCCTATAAAAGGGG GTGGGGGCGCGTTCGTCCTCAGTCGCGATCGAACACTCGAGCCGAG CAGACGTGCCTACGGACCATGGAATTCCCCAGT-3′). The nucleic acid scaffolds that were used to generate the PICs in the open and initial transcribing conformation were designed by modification of the promoter substrate used to form the closed PIC. 11 bp or 19 bp mismatch sequences composed of poly-T were introduced to mimic the different sizes of initiation bubbles in the open and initial transcribing conformation during promoter opening, respectively (Fig. 1, top). Though the mismatch bubble stops at −9 in the open conformation, PIC assembly further unzips two extra base pairs at the upstream end of the bubble. Mismatch bubbles that stop at −11 and −7 were tested in the ITC reconstruction. The latter scaffold did not generate a stable complex, indicating that the GTFs can open up two, but not four, base pairs in the context of the core PIC. An RNA–DNA duplex beyond 7 bp or 12 bp has been proposed to be the trigger for TFIIB release and promoter escape[8,10,18,46,53]. Thus, we introduced a 6 nucleotide (nt) RNA primer complementary to the template strand, starting at the TSS used in our study (non-template 2, 5′- GAAGGGCGCCTATAAAAGGGGGTGGGGGCGCGAAT TTTTTTTTTCGCGATCGAACACTCGAGCCGAGCAGACGTGCCTACGGA CCATGGAATTCCCCAGT-3′; non-template-3, 5′-GAAGGGCGCCTATA AAAGGGGGTGGGGGCGTTTTTTTTTTTTTTTTTTTTCGAACACTCGA GCCGAGCAGACGTGCCTACGGACCATGGAATTCCCCAGT-3′; RNA, 5′-AGUCGC-3′). A biotin tag was engineered at the 5′ end of the template strand (Integrated DNA Technologies). The duplexed DNA was generated by annealing the template strand with equimolar amounts of single-stranded non-template DNA at a final concentration of 50 μM in water. The annealing reaction was carried out at 100 °C for 5 min and gradually cooled down to room temperature within 2 h.

PIC in the closed conformation was assembled as initially described[11], except for an additional incubation of TFIIS at a final concentration of 200 nM with the purified PIC before application to the EM grid[14]. PICs in the open and initial transcribing conformation were assembled in a similar fashion, but using the appropriate nucleic acid scaffolds instead of DNA duplex. Purified PIC complexes were crosslinked after elution by incubation with glutaraldehyde at a final concentration of 0.05%, on ice and under very low illumination conditions, for 5 min, then immediately used for cryo-EM sample preparation.

Electron microscopy. Preparation of PIC samples for cryo-EM observation was carried out using 400 mesh C-flat grids containing 4 μm holes with 4 μm spacing (Protochips). A thin carbon film was floated onto the grid before it was plasma cleaned for 5 s using a Solarus plasma cleaner (Gatan) under a 75% argon/25% oxygen gas mix, immediately before sample deposition. An aliquot (3 μl) of the purified sample (~100 nM) was placed onto the grid and loaded into a Vitrobot (FEI) operated at 100% humidity and 4 °C. The sample was allowed to absorb for 5 min (under low illumination conditions), then was blotted for 4 s and immediate plunged into liquid ethane. The frozen grids were stored in liquid nitrogen until loaded into a Titan electron microscope (FEI) operating at 300 keV using a 626 single-tilt cryotransfer system (Gatan). Data were acquired with a K2 summit direct electron detector (Gatan) operating in counting mode at a calibrated magnification of 37,879 × (1.31 Å per pixel), using a range of defocus values (from −2 to −4 μm). Between 855 and 1334 movie series for each of the cryo-EM data sets were collected using the MSI-T application of the Leginon data collection software[54]. 30-frame exposures were taken at 0.3 s per frame (9 s total exposure time), using a dose rate of 8 e^- per pixel per second, corresponding to a total dose of 42 e^- $Å^{-2}$ per movie series.

Image processing. Movie frames were aligned using MotionCorr to correct for specimen motion[55]. The average of the aligned frames was used for data preprocessing within the Appion processing environment[56]. Particles were automatically selected from the aligned micrographs using a difference of Gaussians (DoG) particle picker[57]. For CC, OC, ITC and ITC(−IIS), 245,501, 328,720, 366,145, and 434,396 particles were picked from 855, 966, 1099, and 1,344 micrographs, respectively (Extended Data Figs 2–4). The contract transfer function (CTF) of each micrograph was estimated using the CTFFind3 program[58] during data collection. All three-dimensional (3D) classification and refinement steps were performed within RELION[59] (version 1.4-beta).

Three-dimensional reconstruction. For the CC, the initial set of 245,501 particles was subjected to an initial 3D classification, using the negative stain reconstruction of a similar complex[11] (in which the nucleic acid is not visible and TFIIS is absent) low-pass filtered to 60 Å as the initial reference (Extended Data Fig. 2c). Because of the use of such reference, the presence of density for the DNA and for TFIIS in the reconstruction can be used as clear signs that features are not arising owing to noise alignment to the references. One out of five classes in this classification, corresponding to 34,728 particles, was indicative of well-preserved holo-complex with sharp structural feature and was selected for further processing. This particle set was then subject to 3D auto-refinement within RELION, resulting in a reconstruction of the CC at an overall resolution of 7.2 Å (Extended Data Fig. 2b). All resolutions reported herein correspond to the gold-standard Fourier shell correlation (FSC) using the 0.143 criterion[60]. Local resolution estimation indicated that the density for TFIIH core (the kinase module is too flexible to be visualized) was at lower resolution than the PIC core containing the rest of the proteins, probably owing to conformational heterogeneity (Extended Data Fig. 2c). Indeed, further 3D classification within RELION successfully separated the data set into two, resulting in two distinct locations of the TFIIH relative to the PIC core (Extended Data Fig. 2c). This flexibility of a major part of the complex would also likely have contributed to the limited resolution of the rest of the CC reconstruction. Therefore, a soft mask was applied around the PIC core density (that is, excluding TFIIH) during further 3D refinement within RELION, thus effectively excluding the contribution of the TFIIH signal from the final alignment rounds. This procedure resulted in an improved reconstruction of the PIC core, with an overall resolution of 5.4 Å (Extended Data Fig. 2b, c).

For the OC, the initial set of 328,720 particles was subjected to an initial 3D classification, similar to the above CC (Extended Data Fig. 3c). One out of five classes in this classification, corresponding to 59,271 particles, was indicative of a well-preserved holo-complex and was selected for further processing. This particle set was then subject to 3D auto-refinement within RELION, resulting in a reconstruction of the OC at 8.6 Å resolution (Extended Data Fig. 3b). Further 3D classification within RELION also successfully separate the data set into two, resulting in two distinct locations of TFIIH relative to the PIC core (Extended Data Fig. 3c). However, the movement trajectory of TFIIH found in the OC was different from the one found in the CC. The same initial data set was also used to derive the reconstruction of the PIC core using the negative stain reconstruction of a similar complex[11] (in which the nucleic acid is not visible and TFIIS is absent) low-pass filtered to 60 Å as the initial reference (Extended Data Fig. 4c). One out of five classes in this classification, corresponding to 79,849 particles, was indicative of well-preserved full core complex and was selected for further processing. This particle set was then subject to 3D auto-refinement followed by particle polishing procedure within RELION[61] in order to correct for individual particle motion and beam-induced radiation damage of the sample. Dose-dependent relative B-factor was estimated per frame and applied during the particle polishing step (Extended Data Fig. 4a). This resulted in a reconstruction of the OC in the absence of TFIIH at an overall resolution of 3.9 Å (Extended Data Fig. 4b). Similar procedures were taken for processing the data set for the ITC and ITC(−IIS) (Extended Data Figs 3, 4).

For the TFIIH core complex, three data sets corresponding to the OC, ITC and ITC(−IIS) were merged together following the initial 3D classification within each data set giving rise to the holo-PIC (Extended Data Fig. 6a). A similar approach to that used for deriving the PIC core within the CC was taken by applying a soft mask around the TFIIH core density during further 3D refinement within RELION, effectively excluding the contribution of core PIC signal during the alignment. This procedure resulted in an improved reconstruction of the TFIIH core, with clear observation of the major and minor grooves of the bound duplex DNA (Fig. 3f and Extended Data Fig. 6a).

Local resolution calculation was performed for all reconstructions using the 'blocres' function in the Bsoft package[62,63] (Extended Data Figs 2–4, 6a). The final density maps were automatically sharpened using the post-processing program within RELION and then filtered according to local resolution using the 'blocfilt' function within Bsoft. Volume segmentation, automatic rigid-body docking, figure and movie generation were performed using UCSF Chimera[64].

Building initial PIC atomic models. To interpret the structural mechanism in more detail, four models of the PIC (CC, OC, ITC and ITC(−IIS)) in different functional states were built. We first generated an initial homology model of human RNA polymerase II (Pol II) in complex with TFIIB using MODELLER[65,66]. We used the sequences of the twelve Pol II subunits (RPB1 to RPB12) and TFIIB without the C-terminal cyclin fold in the crystal structure of yeast *Saccharomyces cerevisiae* Pol II–TFIIB complex (PDB: 4BBR)[10] as a template. This homology model for Pol

© 2016 Macmillan Publishers Limited. All rights reserved

II–TFIIB was placed as a rigid body into the density map using UCSF Chimera[64]. The RPB4 and RPB7 subunits were then replaced by the X-ray structure of human RPB4 and RPB7 heterodimer (PDB: 2C35)[67]. Analysis of the fit showed that the B-linker helix of yeast TFIIB overlapped with the DNA in the cryo-EM reconstruction and that it required to be retraced as a loop region.

The crystal structure of TBP, TFIIB and upstream core promoter DNA duplex (PDB: 1C9B)[38] was placed and rigid-body fitted into the map by aligning TFIIB to the TFIIB's N-terminal cyclin-fold domain. We then extended the model to include TFIIA, by superimposing TBP on the TBP–TFIIA–DNA crystal structure (PDB: 1NVP)[21].

The crystal structure of the human TFIIF dimerization domain (PDB: 1F3U)[68] was docked into the density ascribed to TFIIF by the lobe domain of Pol II. Our rigid body fitting of crystal structures results in a small clash between the RPB2 lobe and the RAP74 C-terminal α1 helix, a region shown to be important in both transcription initiation and elongation. RAP74 α1 helix was rigid-body adjusted to better fit the density and extra residues (residues 167–180) in the C terminus were included based on the results of secondary structure prediction[69,70]. In addition to interacting with Pol II, it is evident from the cryo-EM density that TFIIF directly contacts the BREd DNA. We previously proposed[11] that this region corresponds to the C-terminal winged-helix (WH) domain of RAP30 (PDB: 1BBY)[71], and we can now dock it with high accuracy into our improved map. TFIIF also contributes to the extra density seen linking the C-terminal WH domain and the dimerization domain near the protrusion of Pol II right below the central cleft. We built the path of RAP30, N-terminal to the WH domain on the basis of secondary structure prediction[69,70] and fitted it into the electron microscopy density. Two different protein secondary prediction servers (PSIPRED and SABLE) were used for cross-validation purposes[69,70].

To model the human TFIIE WH domains, residues 11–101, corresponding to the TFIIEα WH, were aligned with *Sulfolobus solfataricus* TFIIE WH residues 7–88 (PDB: 1Q1H)[72]. TFIIEβ WH2 residues 150–207 were aligned with *Sulfolobus tokodaii* STO12A residues 25–106 (PDB: 2D1H)[73]. The human ZR domain, residues 125–164 (PDB: 1VD4)[74], was docked near the base of the stalk domain of Pol II. The human TFIIEβ WH1 residues 75–146 (PDB: 1D8J)[75] were fitted into the region proximal to TFIIF RAP30 WH domain. The individual TFIIE WH domain structures were oriented relative to each other by rigid-body fitting into the electron microscopy density. The TFIIEα residue segments 102–124 and 165–204 and the TFIIEβ residues 205–239 were modelled on the basis of the observed electron microscopy density and consensus secondary structure prediction[69,70]. Missing loops and the linker between the WH domains of TFIIE were built with the program Modeller[65,66].

The X-ray structure of human TFIIS domain II (PDB: 3NDQ) and NMR structure of the TFIIS zinc-finger motif (PDB: 1TFI)[76] were combined and rigid-body fit into the density for the core of Pol II. The linker between domain II and the zinc-finger motif was constructed from the yeast TFIIS structure (PDB: 1Y1V)[26] by homology modelling and fit into the density.

To model TFIIH, we used the crystal structure of the *Archaeoglobus fulgidus* XPB N-terminal domain (PDB: 2FZ4)[77] as a template to build the corresponding N-terminal portion of human XPB. The C-terminal domain of human XPB (PDB: 4ERN)[78] was then directly used to fit into our density map as a rigid body before connecting with the N-terminal part. We generated a homology model for the XPD on the basis of the *Thermoplasma acidophilum* XPD structure (PDB: 2VSF)[79]. The N-terminal helix was added by using *Sulfolobus acidocaldarius* XPD (PDB: 3CRV)[80]. The X-ray structure for Ssl1/P44 from *Saccharomyces cerevisiae* (PDB: 4WFQ)[81] was used as a template to construct the human P44 structure. We used the crystal structure of a minimal complex between Tfb5, the yeast orthologue of P8, and the C-terminal domain of Tfb2, the yeast P52 subunit of TFIIH (PDB: 3DOM)[82], as a template to build the human P52/P8 structure. The human structure of P52/P8 was then docked into a corresponding density adjacent to the C-terminal half of XPB. In order to remove the clashes between the N-terminal helix of P52 and XPB, we rigid-body adjusted the position of this helix to better fit the density. The structure of the N terminus of p52/Tfb2 is predicted to be similar to the HEAT repeats of human transportin 3. We used the human X-ray structure of transportin 4 (PDB: 4C0O)[83] as a template to model the human P52 N-terminal heat repeat. The amino acid sequence of human P52 was aligned with that of human transportin 4, initially with the PROMALS3D multiple sequences and structure alignment server[84], and then adjusted manually. The human p44 VWA domain was modelled using the X-ray structure of the p34 subunit of the TFIIH complex from the eukaryotic thermophilic fungus *Chaetomium thermophilum* (PDB: 4PN7)[85].

The upstream core promoter DNA elements within the OC, ITC and ITC-TFIIS complexes remained bound by TBP, TFIIA and TFIIB, in the same configuration as in the CC. The bent DNA model in the CC was generated using the 3D-DART online server and the 3DNA software package[86]. To model downstream DNA, the DNA–RNA hybrid structure from a previous model of the budding yeast elongation complex (PDB: 4A3F, 4A3D and 4A3I)[5] was fit into the density and adjusted. The downstream DNA was extended by adding duplex DNA in each model. Individual nucleotides in the DNA bubble were adjusted to best represent the observed electron microscopy density in this region of the cryo-EM maps. The sequences of the DNA substrates (in the electron microscopy experiments and the simulation) are shown in Fig. 1a. These initial models (for core CC, OC, ITC, ITC(−IIS), and for TFIIH) were subjected to electron microscopy flexible fitting and refinement of atomic positions as described in the next section.

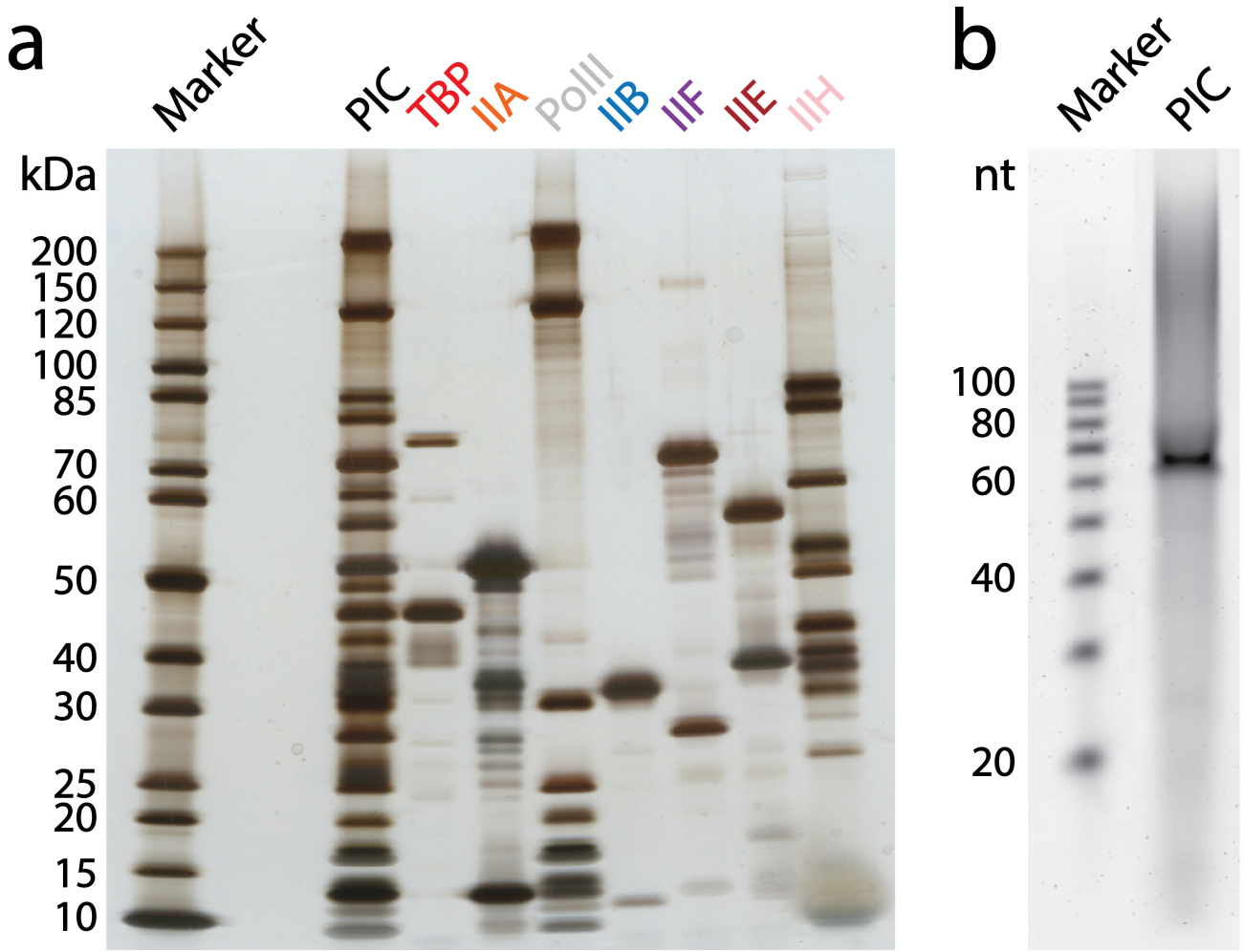
Molecular dynamics flexible fitting and refinement in Phenix. To better capture the conformational changes revealed by the cryo-EM maps in different states, a molecular dynamics flexible fitting procedure (MDFF) was used to flexibly fit the models into the EM density[87]. MDFF is a method wherein external forces proportional to the electron microscopy density gradient are applied, biasing the motion of the protein atoms towards the high-density regions of the map. Additional restraining forces were also applied during fitting to prevent structural distortions and preserve the secondary structure elements. Hydrogen atoms, counterions (Na^+) and TIP3P solvent were introduced using the Tleap module in AMBER10[88]. Additionally, 100 mM NaCl concentration was introduced to mimic physiological conditions. Production runs were carried out in the NPT ensemble (1 atm and 300 K) for 10 ns for each of the four models of the PIC complex (denoted OC, ITC, ITC(−IIS) and CC). We employed MDFF in several stages with progressively higher values for the MDFF force scaling factor ξ. These were varied from 0.1 to 0.3. The XPB/DNA and XPD/P44 systems (IIH model) were simulated for 5 ns with a scaling factor of 0.1. We allowed necessary structural adjustments during MDFF fitting. All simulations were performed using the NAMD2.8 code[89] with the AMBER Parm99SB parameter set containing the basic force field for nucleic acids and proteins, as well as the refined parameters for backbone dihedrals for protein (SB) and nucleic acids dihedrals (BSC0)[90] on the Stampede supercomputer at the Texas Advanced Computing Center and the Edison supercomputer at the National Energy Research Scientific Computing Center (NERSC). MDFF flexible fitting was followed-up by refinement of atomic coordinates and atomic displacement parameters (ADP) using the Phenix package[91]. To carry out reciprocal space refinement, the electron microscopy density maps were converted to structure factors. First, we performed ten macro-cycles of real space refinement applying secondary structure restraints. Then we carried out eight macro-cycles of reciprocal space refinement in Phenix with optimization of the map/stereochemistry weight restraints and applying the MDFF-fitted structures as reference models[92]. To minimize potential overfitting, geometric statistics were monitored during the various rounds of refinements. Finally, the resulting models were validated using MOLPROBITY[93] (see data statistics in Extended Data Table 1). After refinement had converged, the models for the core CC, OC, ITC and ITC(−IIS) and the TFIIH subunits (XPB, XPD, P44, P34, and P52/P8) were separately rigid-body fitted into the respective maps of the holo-complexes (that is, with TFIIH) and the models combined to assemble the full PIC-TFIIH models.

50. Revyakin, A. *et al.* Transcription initiation by human RNA polymerase II visualized at single-molecule resolution. *Genes Dev.* **26,** 1691–1702 (2012).
51. Kassube, S. A., Jinek, M., Fang, J., Tsutakawa, S. & Nogales, E. Structural mimicry in transcription regulation of human RNA polymerase II by the DNA helicase RECQL5. *Nat. Struct. Mol. Biol.* **20,** 892–899 (2013).
52. Juven-Gershon, T., Cheng, S. & Kadonaga, J. T. Rational design of a super core promoter that enhances gene expression. *Nat. Methods* **3,** 917–922 (2006).
53. Pal, M., Ponticelli, A. S. & Luse, D. S. The role of the transcription bubble and TFIIB in promoter clearance by RNA polymerase II. *Mol. Cell* **19,** 101–110 (2005).
54. Suloway, C. *et al.* Automated molecular microscopy: the new Leginon system. *J. Struct. Biol.* **151,** 41–60 (2005).
55. Li, X. *et al.* Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat. Methods* **10,** 584–590 (2013).
56. Lander, G. C. *et al.* Appion: an integrated, database-driven pipeline to facilitate EM image processing. *J. Struct. Biol.* **166,** 95–102 (2009).
57. Voss, N. R., Yoshioka, C. K., Radermacher, M., Potter, C. S. & Carragher, B. DoG Picker and TiltPicker: software tools to facilitate particle selection in single particle electron microscopy. *J. Struct. Biol.* **166,** 205–213 (2009).
58. Mindell, J. A. & Grigorieff, N. Accurate determination of local defocus and specimen tilt in electron microscopy. *J. Struct. Biol.* **142,** 334–347 (2003).
59. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180,** 519–530 (2012).
60. Henderson, R. *et al.* Outcome of the first electron microscopy validation task force meeting. *Structure* **20,** 205–214 (2012).
61. Scheres, S. H. Beam-induced motion correction for sub-megadalton cryo-EM particles. *eLife* **3,** e03665 (2014).
62. Heymann, J. B. & Belnap, D. M. Bsoft: image processing and molecular modeling for electron microscopy. *J. Struct. Biol.* **157,** 3–18 (2007).

© 2016 Macmillan Publishers Limited. All rights reserved

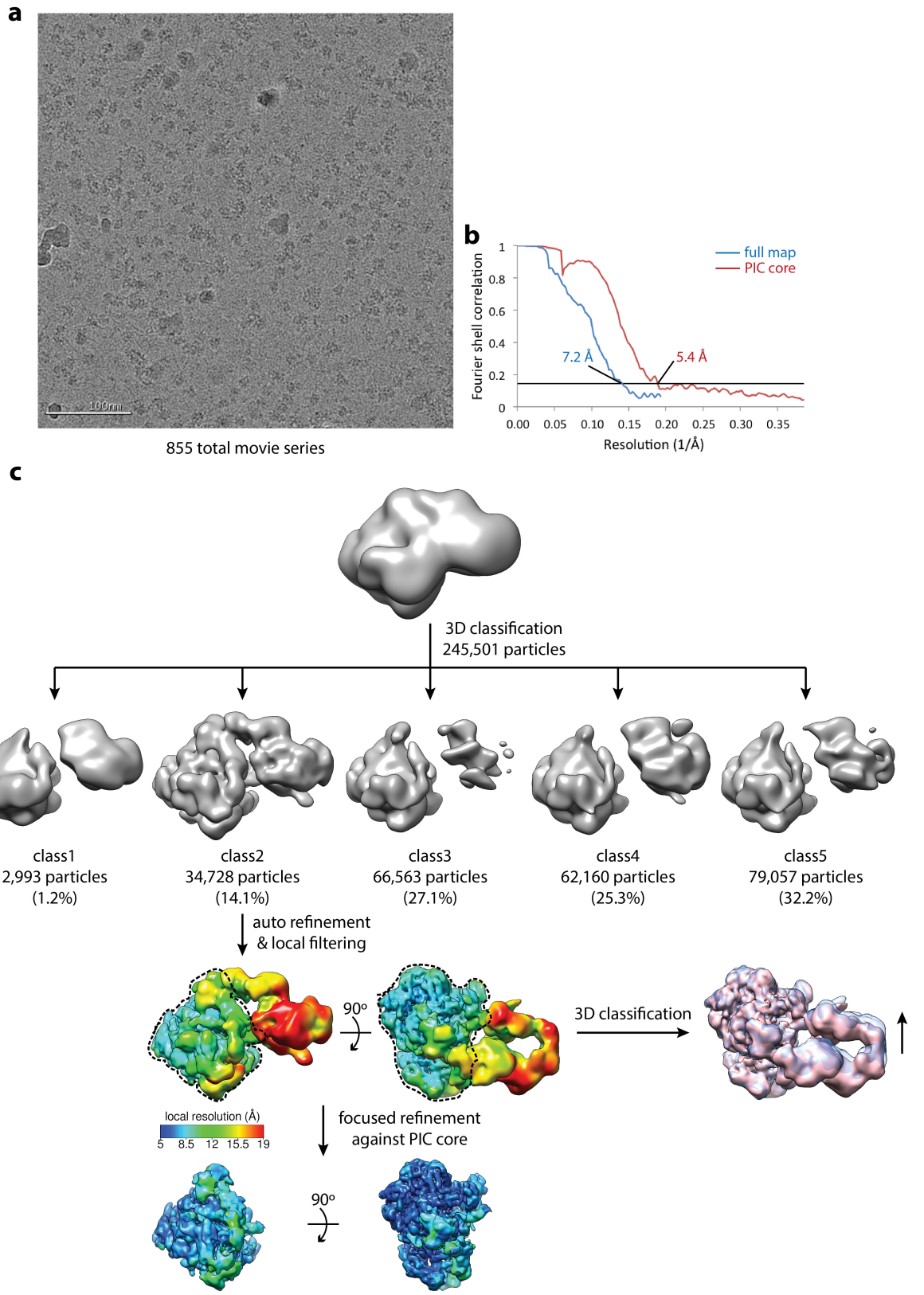
63. Cardone, G., Heymann, J. B. & Steven, A. C. One number does not fit all: mapping local variations in resolution in cryo-EM reconstructions. *J. Struct. Biol.* **184,** 226–236 (2013).
64. Goddard, T. D., Huang, C. C. & Ferrin, T. E. Visualizing density maps with UCSF Chimera. *J. Struct. Biol.* **157,** 281–287 (2007).
65. Sali, A. Comparative protein modeling by satisfaction of spatial restraints. *Mol. Med. Today* **1,** 270–277 (1995).
66. Martí-Renom, M. A. *et al.* Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29,** 291–325 (2000).
67. Meka, H., Werner, F., Cordell, S. C., Onesti, S. & Brick, P. Crystal structure and RNA binding of the Rpb4/Rpb7 subunits of human RNA polymerase II. *Nucleic Acids Res.* **33,** 6435–6444 (2005).
68. Gaiser, F., Tan, S. & Richmond, T. J. Novel dimerization fold of RAP30/RAP74 in human TFIIF at 1.7 Å resolution. *J. Mol. Biol.* **302,** 1119–1127 (2000).
69. Adamczak, R., Porollo, A. & Meller, J. Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins* **56,** 753–767 (2004).
70. Buchan, D. W. A., Minneci, F., Nugent, T. C. O., Bryson, K. & Jones, D. T. Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res.* **41,** W349–W357 (2013).
71. Groft, C. M., Uljon, S. N., Wang, R. & Werner, M. H. Structural homology between the Rap30 DNA-binding domain and linker histone H5: implications for preinitiation complex assembly. *Proc. Natl Acad. Sci. USA* **95,** 9117–9122 (1998).
72. Meinhart, A., Blobel, J. & Cramer, P. An extended winged helix domain in general transcription factor E/IIE alpha. *J. Biol. Chem.* **278,** 48267–48274 (2003).
73. Shinkai, A. *et al.* The putative DNA-binding protein Sto12a from the thermoacidophilic archaeon Sulfolobus tokodaii contains intrachain and interchain disulfide bonds. *J. Mol. Biol.* **372,** 1293–1304 (2007).
74. Okuda, M. *et al.* A novel zinc finger structure in the large subunit of human general transcription factor TFIIE. *J. Biol. Chem.* **279,** 51395–51403 (2004).
75. Okuda, M. *et al.* Structure of the central core domain of TFIIEbeta with a novel double-stranded DNA-binding surface. *EMBO J.* **19,** 1346–1356 (2000).
76. Qian, X. *et al.* Novel zinc finger motif in the basal transcriptional machinery: three-dimensional NMR studies of the nucleic acid binding domain of transcriptional elongation factor TFIIS. *Biochemistry* **32,** 9944–9959 (1993).
77. Fan, L. *et al.* Conserved XPB core structure and motifs for DNA unwinding: implications for pathway selection of transcription or excision repair. *Mol. Cell* **22,** 27–37 (2006).
78. Hilario, E., Li, Y., Nobumori, Y., Liu, X. & Fan, L. Structure of the C-terminal half of human XPB helicase and the impact of the disease-causing mutation XP11BE. *Acta Crystallogr. D* **69,** 237–246 (2013).
79. Wolski, S. C. *et al.* Crystal structure of the FeS cluster-containing nucleotide excision repair helicase XPD. *PLoS Biol.* **6,** e149 (2008).
80. Fan, L. *et al.* XPD helicase structures and activities: insights into the cancer and aging phenotypes from XPD mutations. *Cell* **133,** 789–800 (2008).
81. Kim, J. S. *et al.* Crystal structure of the Rad3/XPD regulatory domain of Ssl1/p44. *J. Biol. Chem.* **290,** 8321–8330 (2015).
82. Kainov, D. E., Vitorino, M., Cavarelli, J., Poterszman, A. & Egly, J. M. Structural basis for group A trichothiodystrophy. *Nat. Struct. Mol. Biol.* **15,** 980–984 (2008).
83. Maertens, G. N. *et al.* Structural basis for nuclear import of splicing factors by human Transportin 3. *Proc. Natl Acad. Sci. USA* **111,** 2728–2733 (2014).
84. Pei, J., Kim, B. H., Tang, M. & Grishin, N. V. PROMALS web server for accurate multiple protein sequence alignments. *Nucleic Acids Res.* **35,** W649–W652 (2007).
85. Schmitt, D. R., Kuper, J., Elias, A. & Kisker, C. The structure of the TFIIH p34 subunit reveals a von Willebrand factor A like fold. *PLoS One* **9,** e102389 (2014).
86. Lu, X. J. & Olson, W. K. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.* **31,** 5108–5121 (2003).
87. Trabuco, L. G., Villa, E., Mitra, K., Frank, J. & Schulten, K. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* **16,** 673–683 (2008).
88. Cornell, W. D. *et al.* A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117,** 5179–5197 (1995).
89. Kalé, L. *et al.* NAMD2: Greater scalability for parallel molecular dynamics. *J. Comput. Phys.* **151,** 283–312 (1999).
90. Hornak, V. *et al.* Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **65,** 712–725 (2006).
91. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
92. Adams, P. D. *et al.* The Phenix software for automated determination of macromolecular structures. *Methods* **55,** 94–106 (2011).
93. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66,** 12–21 (2010).
94. Okuda, M. *et al.* Structural insight into the TFIIE-TFIIH interaction: TFIIE and p53 share the binding region on TFIIH. *EMBO J.* **27,** 1161–1171 (2008).

© 2016 Macmillan Publishers Limited. All rights reserved



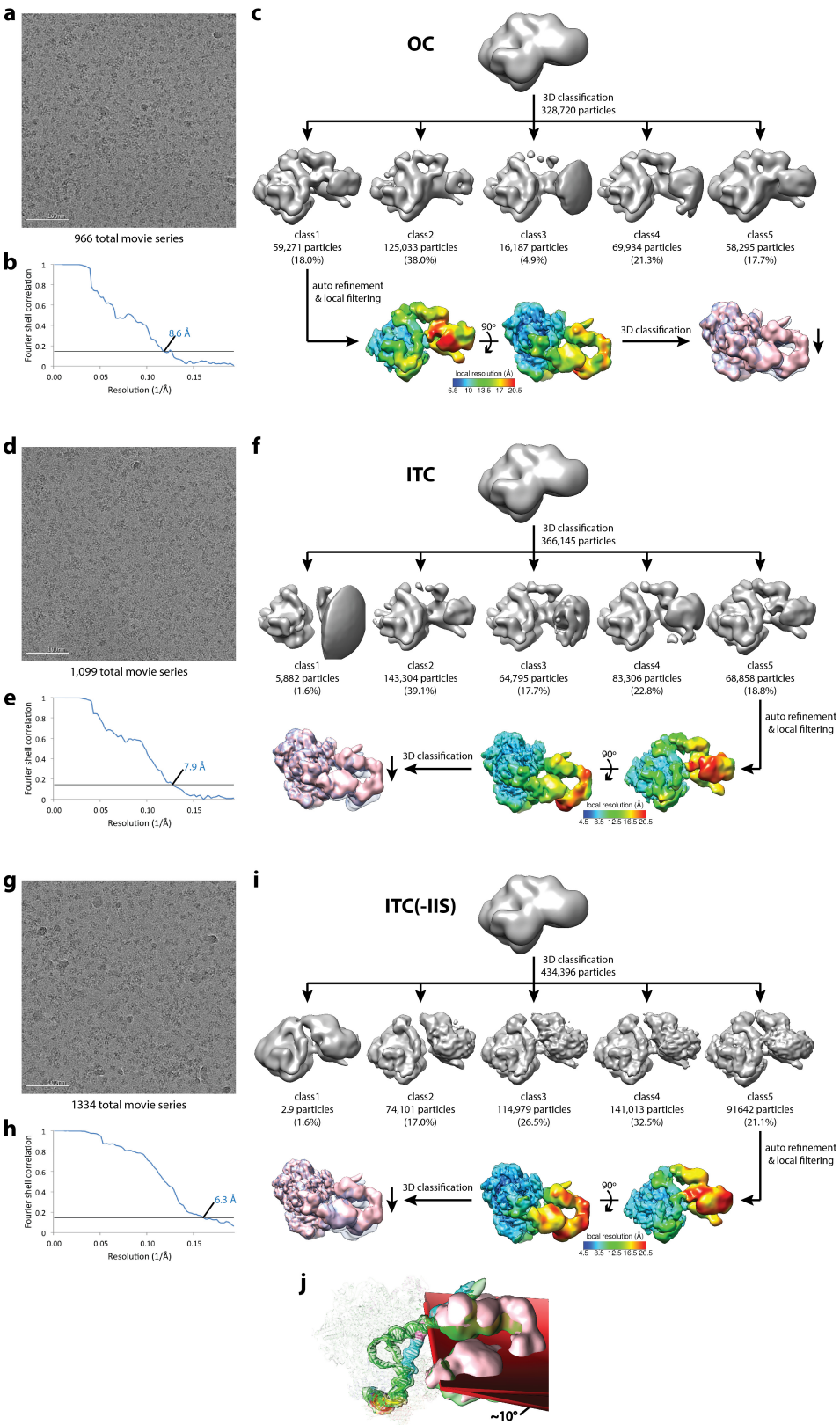
Extended Data Figure 1 | Purification and activity of human PIC. a, SDS–PAGE (4–12% gradient gel followed by silver staining) of purified transcription factors and assembled PIC on a promoter DNA. **b**, The purified PIC in **a** was ran on a 10% TBE-urea gel after supplying ribonucleoside triphosphates in a run-off reaction. kDa, kilodaltons.

© 2016 Macmillan Publishers Limited. All rights reserved



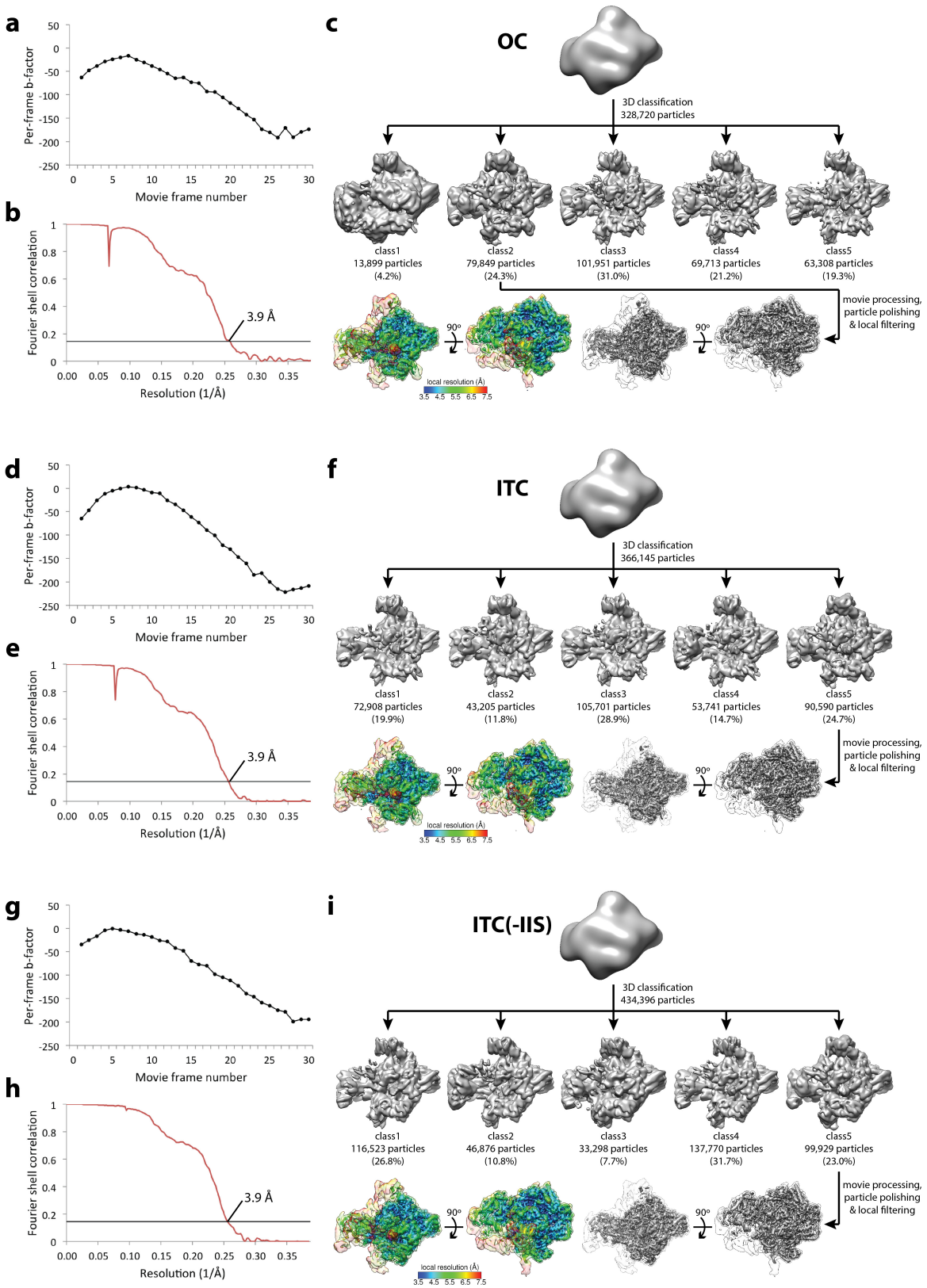
Extended Data Figure 2 | Cryo-EM of the human CC. **a**, Representative raw micrograph. **b**, FSC curve and estimated resolution using the 0.143 criteria following the gold-standard procedure implemented in RELION for both the holo-complex and the PIC core. **c**, Refinement strategy for the holo-complex (see Methods). The local resolution estimation shows flexibility for TFIIH. Further 3D classification revealed the range of motion of TFIIH within the complex (pink and sky blue densities). Focused refinement on the PIC core (masking out TFIIH) improved alignment accuracy and improved the resolution for the core complex.

© 2016 Macmillan Publishers Limited. All rights reserved



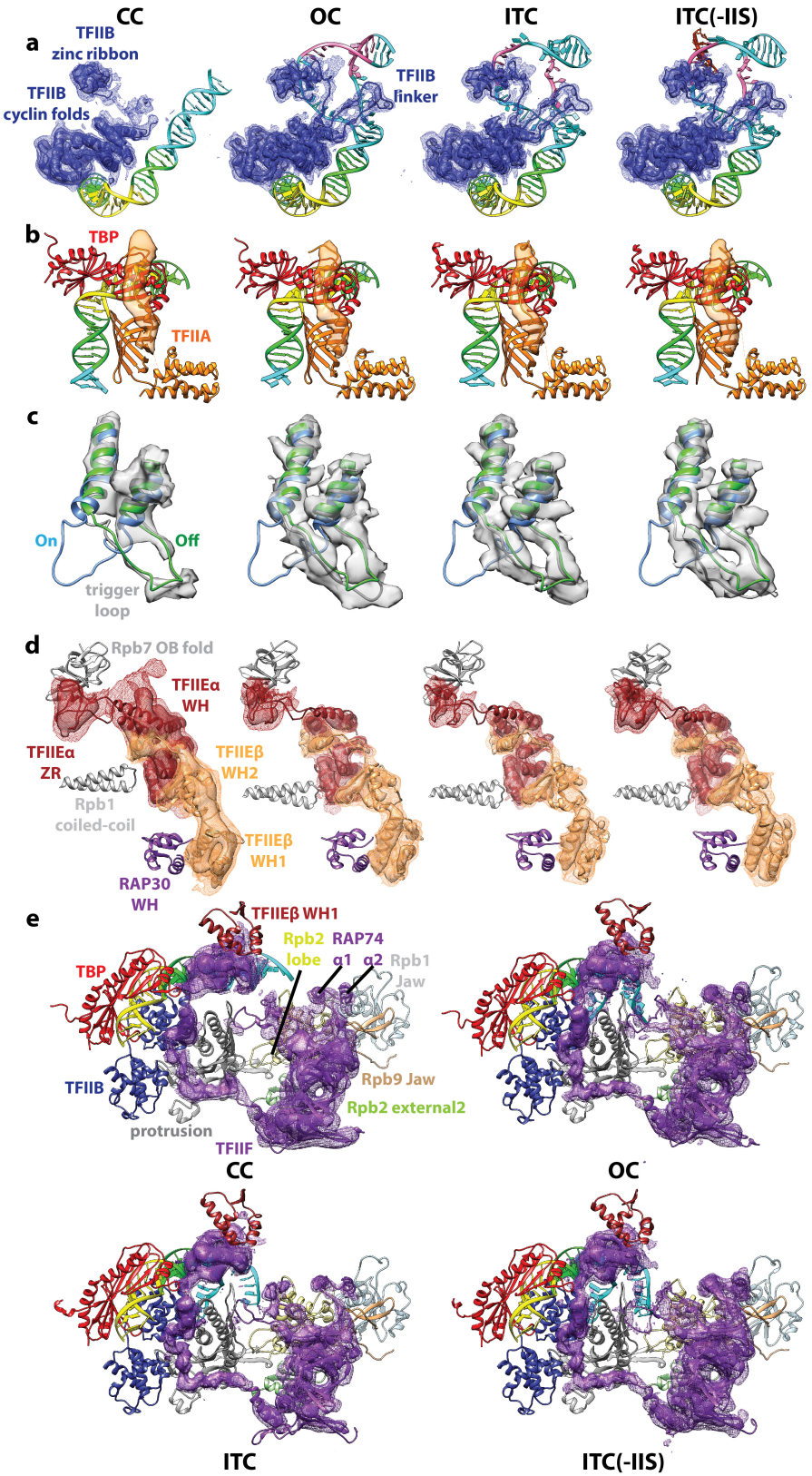
Extended Data Figure 3 | Cryo-EM of the human OC/ITC/ITC(−IIS) complexes. a, **d**, **g**, Representative raw micrograph. **b**, **e**, **h**, FSC curve and estimated resolution using the 0.143 criteria following the gold-standard procedure implemented in RELION. **c**, **f**, **i**, Refinement strategy for the holo-complex (see Methods). The local resolution estimation shows flexibility for TFIIH. Further 3D classification revealed the range of motion of TFIIH within the complex (pink and blue densities). Notice that the direction and range of motion for these three states is similar. Compared to the CC, the pink densities are approximately the same, but the direction of motion from pink to blue has changed. **j**, Movement of TFIIH with respect to the core PIC from the CC to OC state.

© 2016 Macmillan Publishers Limited. All rights reserved



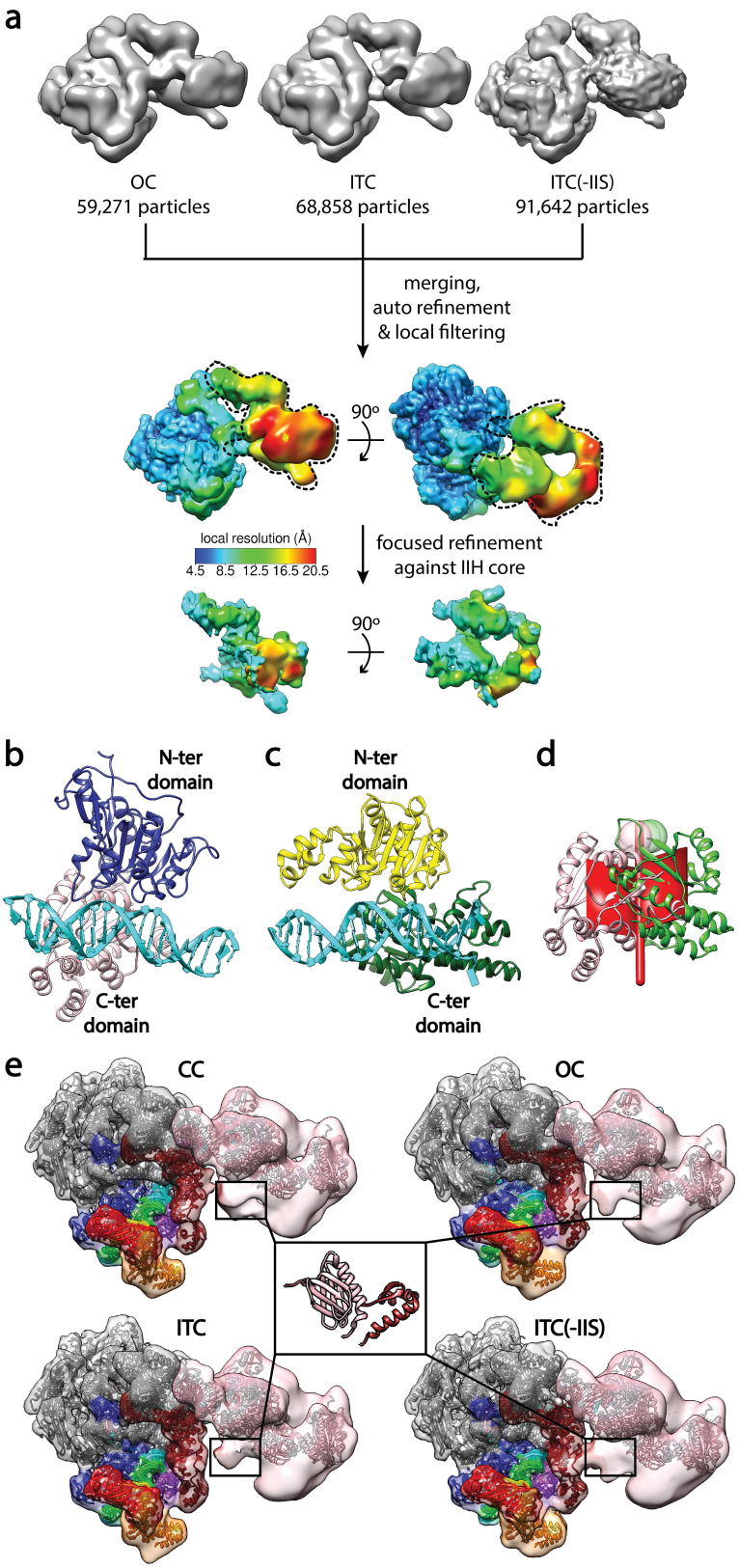
Extended Data Figure 4 | Cryo-EM of the core human OC/ITC/ITC(−IIS) complexes. a, **d**, **g**, Dose-dependent B-factor plot from RELION. **b**, **e**, **h**, FSC curve and estimated resolution using the 0.143 criteria following the gold standard procedure implemented in RELION. **c**, **f**, **i**, Refinement strategy (see Methods). The local resolution estimation and final density map (filtered according to this local resolution) is shown for two different views and at two different thresholds (the higher threshold allows better visualization of the highest resolution features for the more stable elements).

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 5 | Segmented cryo-EM densities and models for different regions of interest within the human core PIC in the CC, OC, ITC and ITC(−IIS) states. a, TFIIB and its interaction with DNA. **b**, Extension of TFIIA stabilized by interaction with TBP. **c**, Trigger loop in Pol II, viewed in an 'off' state by comparison with crystallographic structures of yeast Pol II (2NVZ[25] in blue and 1Y1V[26] in green). **d**, TFIIE and its interacting partners with the PIC. **e**, TFIIF and its interacting partners with the PIC.

© 2016 Macmillan Publishers Limited. All rights reserved

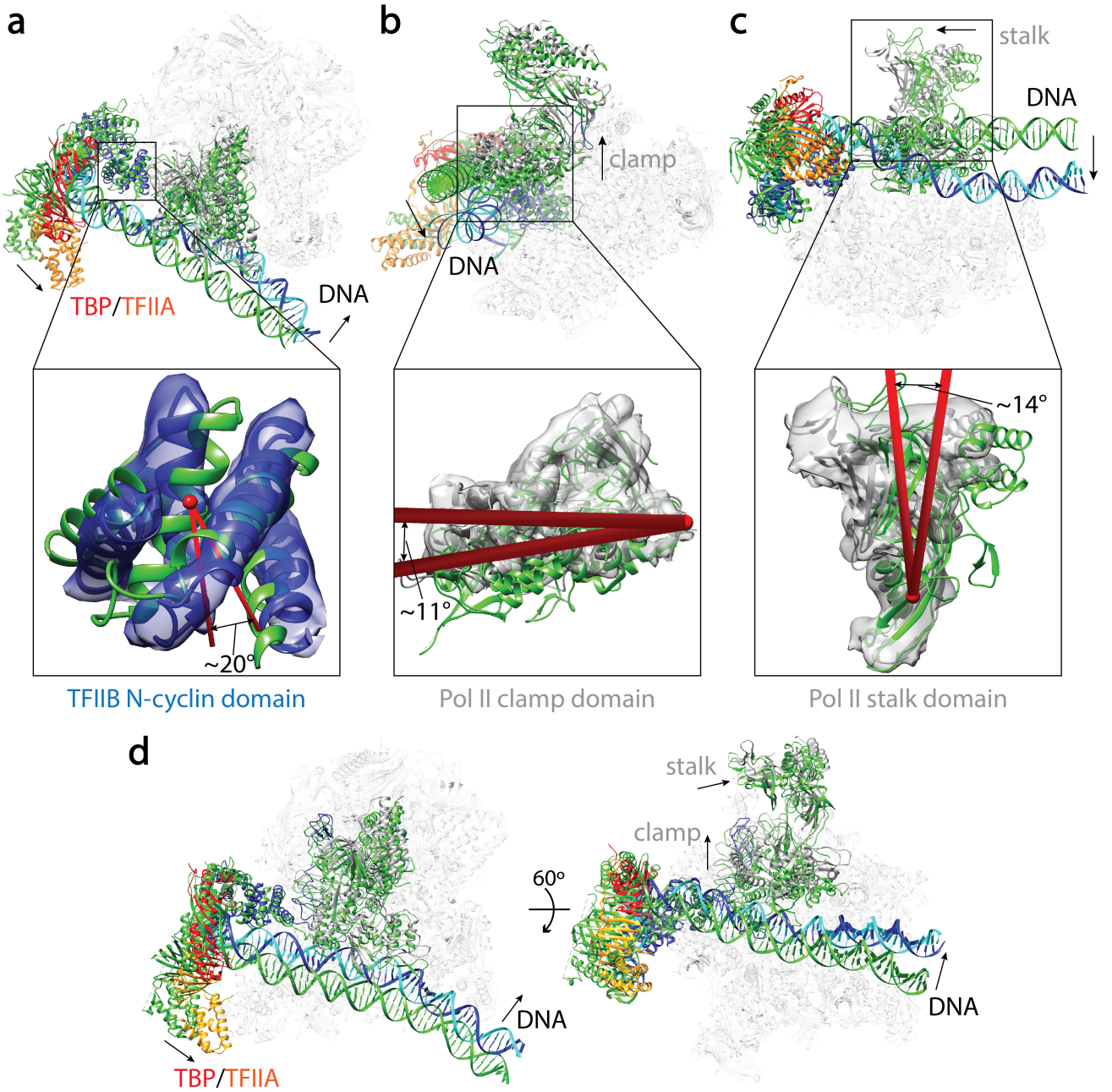


Extended Data Figure 6 | See next page for caption.

© 2016 Macmillan Publishers Limited. All rights reserved

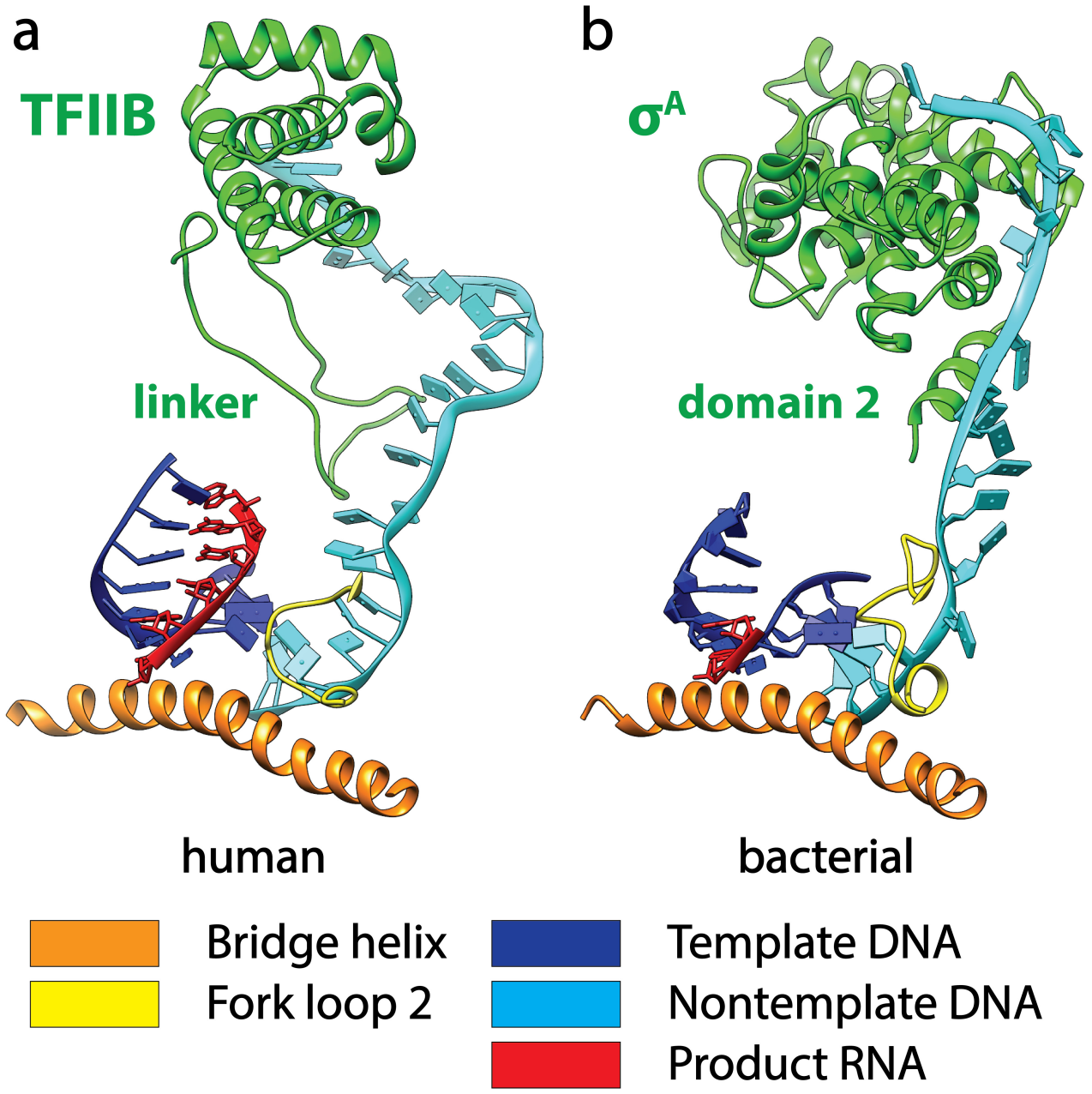
Extended Data Figure 6 | Cryo-EM reconstruction and interpretation of human TFIIH. a, Refinement strategy (see Methods) for the seven subunit TFIIH core complex (the CAK subcomplex is too flexible to be visualized following averaging). The local resolution estimation for the PIC structure obtained by combining the OC, ITC and ITC(−IIS) data sets shows lower resolution for TFIIH, reflecting its flexible attachment. Focused refinement (using the mask marked by dashed lines) allowed us to improve the resolution for TFIIH. **b**–**d**, Comparison of human XPB and the ssoRad54 ATPase structures and their interaction with DNA. **b**, Human XPB–DNA model. N terminus is shown in navy blue, C terminus in pink and DNA in cyan. **c**, Crystal structure of ssoRad54 ATPase (PDB: 1Z63)[33]. The N terminus is shown in yellow, the C terminus in green and the DNA in cyan. Whereas the N-terminal domains of both proteins and the DNA can be easily superimposed, the C-terminal domains are in very different position. **d**, A 123° rotation would be required to superimpose the C-terminal domains around an axis located on the first residues (top, shown in spheres) that connects to the fixed N-terminal domain. The rotation axis and planes are coloured in red. **e**, Possible location of the TFIIEα–TFIIH/p62 interface. An unassigned density in the region of proximity/contact between TFIIE and TFIIH probably corresponding to the TFIIE–TFIIH interface is marked with boxes for each reconstruction. A tentative orientation of the NMR structure for a short C-terminal segment of TFIIEα bound to the PHD domain of p62 (PDB: 2RNR)[94] is proposed in the centre for that flexible density region.

© 2016 Macmillan Publishers Limited. All rights reserved



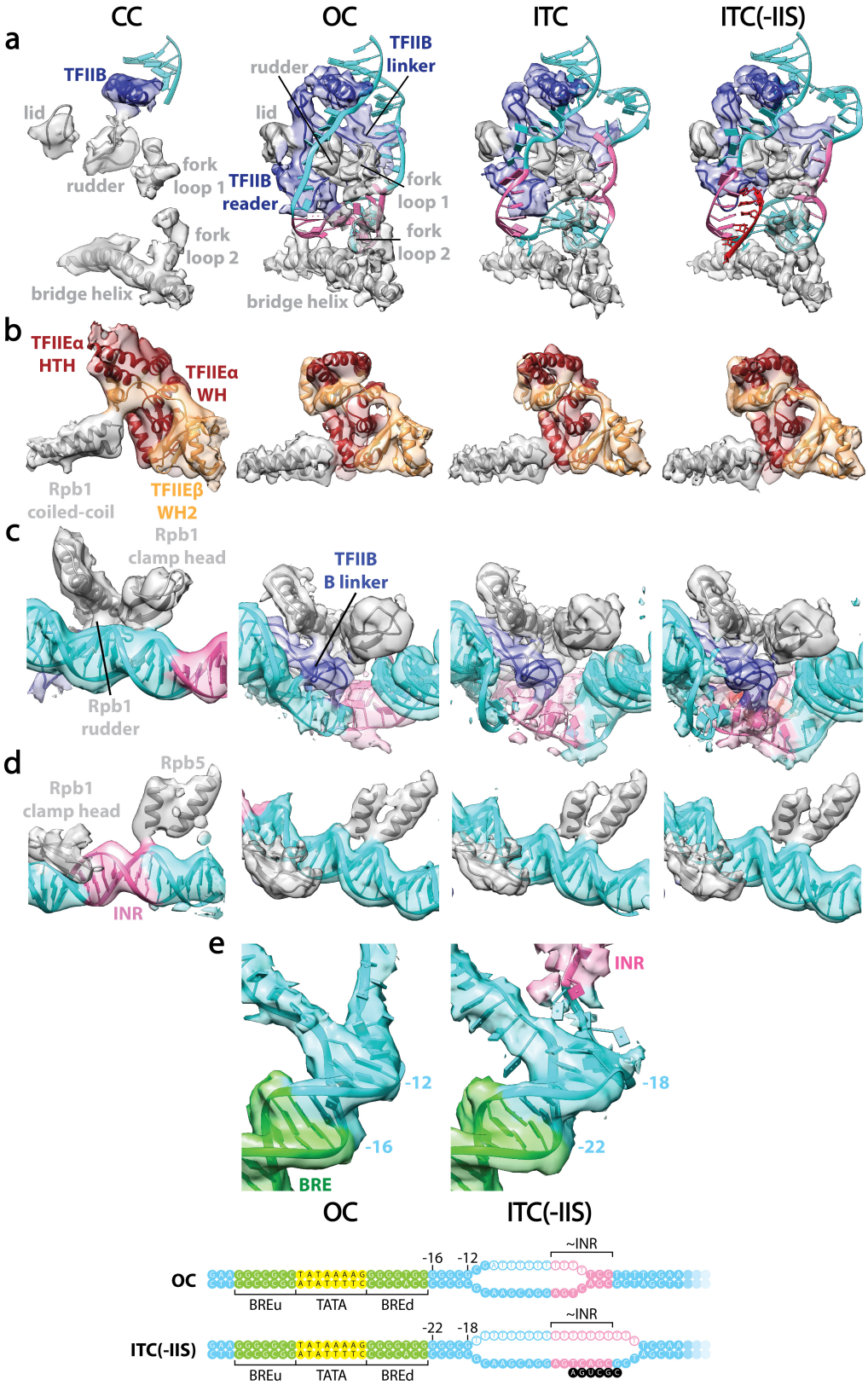
Extended Data Figure 7 | Position of the TBP–IIA–IIB module and mobile elements of Pol II. a–c, Comparison between a 'synthetic' structure of the CC, generated by superimposing human TBP–TFIIA–DNA (PDB: 1NVP)[21], human TBP–TFIIB–DNA (PDB: 2C9B)[38], and, most notably, the yeast Pol II–TFIIB (PDB: 4BBR)[10] using common elements (green), and the human cryo-EM CC model (coloured). A number of elements are rotated between the two: the TBP-IIA-IIB-DNA subcomplex (**a**), the Pol II clamp (**b**), and the Pol II stalk (**c**). The rotation plane and angle are depicted in red. **d**, Comparison, shown in two different views, of a recently reported cryo-EM structure of yeast CC (green)[13] and the human CC in this study. The structures were aligned using the rigid part of Pol II (that is, excluding the clamp and the stalk). The yeast TBP–TFIIA–TFIIB–DNA module and the mobile regions of Pol II (clamp and stalk) are in different relative positions. Whereas the TBP–TFIIA–TFIIB module and the clamp and stalk element resembles those in the 'synthetic' model (**a–c**), the path of the DNA is very different, in that it moves away, rather than towards the Pol II stalk.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 8 | Comparison of eukaryotic and bacterial initiation complexes around the active site. a, **b**, Equivalent close-up views of our ITC(−IIS) model (**a**) and the crystallographic structure of a bacterial initiation complex (4G7O[42]) (**b**). The fork loop 2 is tilted in a very similar manner to that observed in our OC, ITC and ITC(−IIS) structures (see Figs 2b and 5f). Domain 2 within the bacterial σ factor is involved in stabilizing the non-template DNA in a similar manner as the TFIIB linker region in human initiation complex.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 9 | Comparison of selected regions of the human core PIC structures for the CC, OC, ITC and ITC(−IIS) states. **a**, Segmented density and model for TFIIB and Pol II loops critical for stabilization of the transcription bubble. The density for the nucleic acids has been omitted for clarity. **b**, Interaction of the Pol II clamp head with TFIIE. As the clamp closes down during promoter opening (CC to OC transition), the region of contact with TFIIE changes. **c**, Interaction of the Pol II clamp head with DNA near the promoter melting site. **d**, Interaction of RPB5 with DNA. **e**, Opening of two extra base pairs of DNA in the OC scaffold. EM density and the corresponding model near the initiation bubble upstream fork in the OC (left) and ITC(−IIS) (right) structures. Positions of the duplexed DNA downstream of the BRE were labelled relative to the +1 active site in each structure. The cartoon (bottom) shows the two aligned DNA templates for reference.

© 2016 Macmillan Publishers Limited. All rights reserved

Extended Data Table 1 | Statistics after refinement of atomic coordinates with Phenix for the core PIC models (OC, ITC, ITC(−IIS) and CC), and scores from MolProbilty structural analysis

	OC	ITC	ITC without TFIIS	CC
R	0.31	0.30	0.305	0.32
R-free	0.32	0.31	0.306	0.34
RMS angles (˚)	1.41	1.49	1.48	1.10
RMS bonds (Å)	0.010	0.012	0.011	0.005
Ramachandran outliers (%)	2.8	3.0	2.67	2.6
Ramachandran favored (%)	88.4	88.7	88.55	89.9
Clashscore	12.77	12.87	14.10	22.04
Overall MolProbity score	2.42	2.50	2.44	2.54

© 2016 Macmillan Publishers Limited. All rights reserved

LETTER

doi:10.1038/nature17952

An irradiated brown-dwarf companion to an accreting white dwarf

Juan V. Hernández Santisteban[1], Christian Knigge[1], Stuart P. Littlefair[2], Rene P. Breton[1,3], Vikram S. Dhillon[2,4], Boris T. Gänsicke[5], Thomas R. Marsh[5], Magaretha L. Pretorius[6], John Southworth[7] & Peter H. Hauschildt[8]

Interacting compact binary systems provide a natural laboratory in which to study irradiated substellar objects. As the mass-losing secondary (donor) in these systems makes a transition from the stellar to the substellar regime, it is also irradiated by the primary (compact accretor)[1,2]. The internal and external energy fluxes are both expected to be comparable in these objects, providing access to an unexplored irradiation regime. The atmospheric properties of donors are largely unknown[3], but could be modified by the irradiation. To constrain models of donor atmospheres, it is necessary to obtain accurate observational estimates of their physical properties (masses, radii, temperatures and albedos). Here we report the spectroscopic detection and characterization of an irradiated substellar donor in an accreting white-dwarf binary system. Our near-infrared observations allow us to determine a model-independent mass estimate for the donor of 0.055 ± 0.008 solar masses and an average spectral type of L1 ± 1, supporting both theoretical predictions and model-dependent observational constraints that suggest that the donor is a brown dwarf. Our time-resolved data also allow us to estimate the average irradiation-induced temperature difference between the dayside and nightside of the substellar donor (57 kelvin) and the maximum difference between the hottest and coolest parts of its surface (200 kelvin). The observations are well described by a simple geometric reprocessing model with a bolometric (Bond) albedo of less than 0.54 at the 2σ confidence level, consistent with high reprocessing efficiency, but poor lateral heat redistribution in the atmosphere of the brown-dwarf donor[4,5]. These results add to our knowledge of binary evolution, in that the donor has survived the transition from the stellar to the substellar regime, and of substellar atmospheres, in that we have been able to test a regime in which the irradiation and the internal energy of a brown dwarf are comparable.

So far, only a single brown-dwarf donor to an accreting white dwarf has been detected spectroscopically[3], and no empirical information is available regarding the effect of irradiation on the atmosphere of the donor. We have therefore carried out simultaneous optical and near-infrared time-resolved spectroscopy of one such system with the X-Shooter instrument[6] at the Very Large Telescope (European Southern Observatory, ESO). Our target, SDSS (Sloan Digital Sky Survey) J143317.78+101123.3 (J1433 hereafter; ref. 7), is an eclipsing system with a short orbital period ($P_{\rm orb} = 78.1$ min) and a donor mass that is probably well below the hydrogen-burning limit, as estimated from eclipse modelling ($M_2 = (0.0571 \pm 0.0007) M_{\odot}$; refs 8, 9). The exceptionally wide wavelength coverage provided by our data set (0.35–2.5 μm) allows us to confidently dissect the overall spectral energy distribution. This is illustrated in Fig. 1. Disk-formed, double-peaked hydrogen emission lines can be seen across the entire spectral range. The presence of broad Balmer absorption lines between 0.3 μm and 0.5 μm shows that the white dwarf is the strongest emitter in this wave band. However, the spectrum of the donor star is clearly visible in the near-infrared and dominates between 1.0 μm and 2.5 μm.

The clear donor signature in the near-infrared allows us to isolate the spectrum of the donor and determine its properties. We decomposed the average spectrum by fitting the optical region with a white-dwarf atmosphere model[10] and an accretion disk (modelled as a simple power law). Both of these contributions were then extrapolated to the near-infrared and subtracted to retrieve a pure donor spectrum. This donor spectrum was then compared to an empirical brown-dwarf spectral sequence[11]. The best-matching spectral type was found to be SpT = L1 ± 1, which agrees remarkably well with semi-empirical evolutionary predictions[1,2] and with previous narrow J-band estimates[3].

These results imply that the donor in J1433 has successfully undergone the transition from the stellar to the substellar regime. The alternative—that the secondary was born as a brown dwarf—is extremely unlikely. First, extreme-mass-ratio binary systems containing very-low-mass objects ($M \lesssim 0.1 M_{\odot}$) are intrinsically rare[12]. Second, the mass of the accreting white dwarf in J1433 ($M_{\rm WD} = (0.80 \pm 0.07) M_{\odot}$) is exactly in line with the mean white-dwarf mass of known accreting white-dwarf systems going through the standard evolution channel[13], but much higher than the average mass of isolated white dwarfs ($\langle M_{\rm WD} \rangle \approx 0.6 M_{\odot}$; ref. 13) and that of any known primary in a white-dwarf–brown-dwarf binary system[14]. Third, the orbital period of J1433 lies squarely within the 'period minimum spike' that coincides with the stellar-to-substellar transition of the donor star in systems born with main-sequence companions[15]. By contrast, systems born with substellar donors would be expected to populate a wide range of orbital periods below this spike[16]. However, no such systems have been found so far. Indeed, the only accreting white dwarf with a substellar donor and an orbital period well below the period minimum has been shown to be a low-metallicity object in the Galactic halo[17], rather than a system born with a brown-dwarf secondary.

Models of irradiated planets suggest that irradiation can increase photospheric temperatures by an order of magnitude compared to an isolated object, with substantial effects on the radius and atmospheric structure of the planet[18]. Furthermore, a long-standing mismatch between the predicted and observed minimum period for accreting white-dwarf binary systems can be traced to an approximately 10% offset between the donor radii inferred from observations and those predicted by theoretical model atmospheres[19]. However, such large effects are only possible if the irradiating flux is efficiently absorbed in the atmosphere and redistributed across the terminator, that is, from the dayside to the nightside[2,20]. This is necessary because it is the blocking effect of the incoming radiation on the outward heat flux from the interior of the secondary that drives the irradiation-induced swelling of the donor (secondary).

Globally, the bolometric flux emitted by the irradiated donor must always balance the sum of its intrinsic flux and the irradiation flux it

[1]Department of Physics and Astronomy, University of Southampton, Southampton SO17 1BJ, UK. [2]Department of Physics and Astronomy, University of Sheffield, Sheffield S3 7RH, UK. [3]Jodrell Bank Centre for Astrophysics, School of Physics and Astronomy, The University of Manchester, Manchester M13 9PL, UK. [4]Instituto de Astrofísica de Canarias, E-38205 La Laguna, Santa Cruz de Tenerife, Spain. [5]Department of Physics, University of Warwick, Coventry CV4 7AL, UK. [6]Department of Physics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, UK. [7]Astrophysics Group, Keele University, Newcastle ST5 5BG, UK. [8]Hamburger Sternwarte, Gojenbergsweg 112, 21029 Hamburg, Germany.

© 2016 Macmillan Publishers Limited. All rights reserved

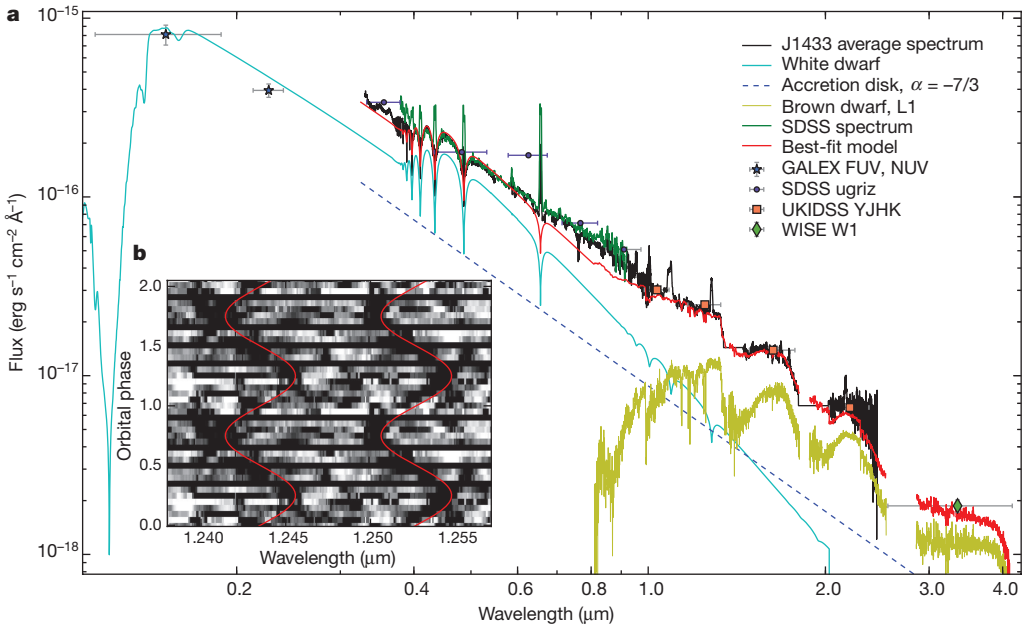


Figure 1 | The average optical and near-infrared spectral energy distribution of SDSS J143317.78+101123.3. a, The overall fit to the spectra (black line) using a white-dwarf atmosphere model (light blue line) and brown-dwarf templates (yellow line) was performed. The contribution from the accretion disk (dark blue dashed line) was described as a simple power law with an index $\alpha = -7/3$. We obtained a best fit (red line) with a spectral type of SpT = L1 ± 1. GALEX FUV, NUV (blue stars), SDSS (green line, purple circle), UKIDSS (orange squares) and WISE (green diamonds) photometry are shown for reference; vertical error bars represent the 1σ confidence intervals and horizontal error bars represent the width of the filter. **b**, Trailed, phase-binned spectra around the K I doublet at 1.2436 μm and 1.2528 μm. These absorption features arise in the atmosphere of the donor star (greyscale), enabling the determination of orbital parameters such as the semi-amplitude $K_2 = 499 \pm 15$ km s^{-1}, the rotational broadening of the donor $v\sin i = 131 \pm 46$ km s^{-1} and the systemic velocity $\gamma = -42 \pm 8$ km s^{-1} (red line).

has absorbed[21]. However, in the absence of redistribution, this balance must be maintained locally in the atmosphere. This should produce pronounced temperature differences between the dayside and nightside on the tidally locked donor. Because the system is seen close to edge-on, such differences should manifest as changes in the apparent spectral type of the donor as a function of orbital phase. We have tested for this by measuring the H_2O 1.3-μm-band index[11]. Water vapour is the most sensitive absorber in irradiated hot Jupiters[22], and water-band strength has been shown to correlate well with spectral type and effective temperature[11]. As shown in Fig. 2b, we detect a spectroscopic

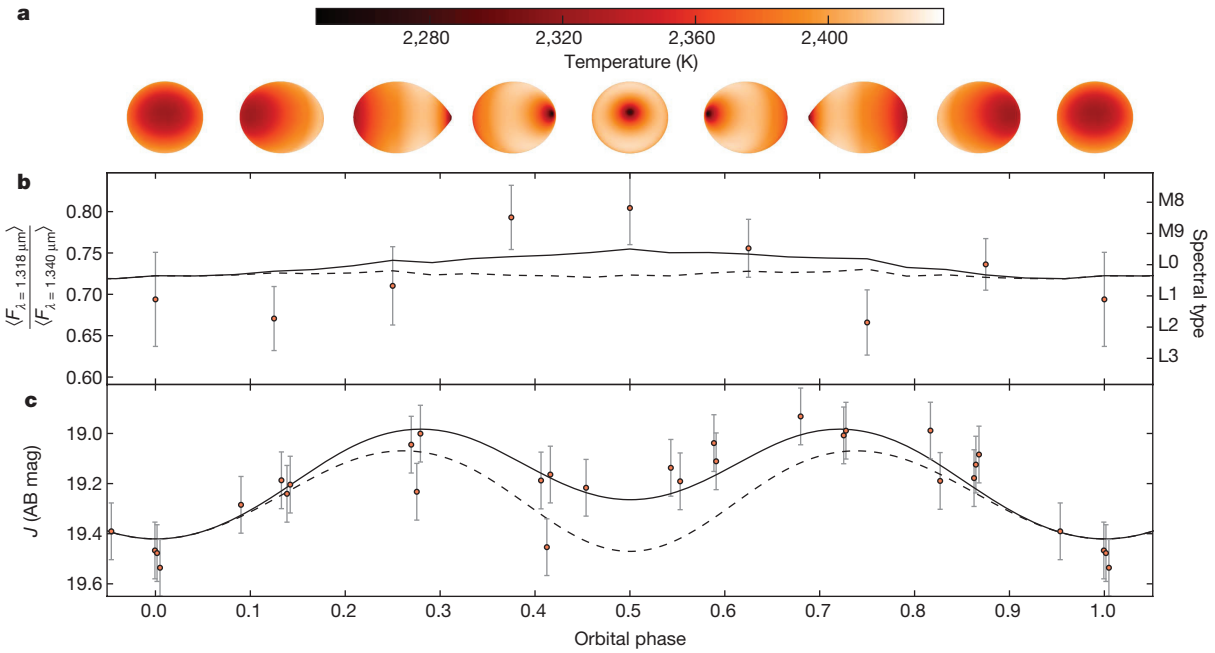


Figure 2 | Ellipsoidal and irradiation effects associated with the substellar donor in SDSS J143317.78+101123.3. We find a difference in spectral type between the dayside and nightside hemispheres of $\Delta\mathrm{SpT} \approx 1$, equivalent to $\Delta T \approx 57$ K with a maximum temperature difference of approximately 200 K. **a**, Temperature distribution of the atmosphere of the donor as a function of phase. **b**, 1.3-μm water-band depth, $\langle F_{\lambda=1.318\,\mu m}\rangle / \langle F_{\lambda=1.340\,\mu m}\rangle$, measurements (circles) with the best-fit model (solid line) and a model with no irradiation source (dashed line) for comparison. **c**, Timed-resolved J-band photometry (circles) after subtracting the contributions from the white dwarf and the accretion disk. Lines represent the same models as in **b**. Error bars in **b** and **c** represent 1σ confidence intervals.

© 2016 Macmillan Publishers Limited. All rights reserved

change of $\Delta\mathrm{SpT} \approx 1$ (that is, M9 to L0), with a maximum at phase of 0.5, precisely where the dayside of the donor is exposed.

On its own, this spectroscopic signal would be only marginally significant, because a model without irradiation can be ruled out at only a 2σ confidence level. However, an independent irradiation signature is provided by the observed broadband flux variations around the orbit, which are shown in Fig. 2c. This figure reveals a double modulation over one orbit, which is produced by the combination of two effects: the distorted shape of the tidally locked, Roche-lobe-filling donor star (ellipsoidal variations) and the temperature difference between its dayside and nightside (irradiation effect). To quantify the irradiation effect, we used the binary light curve synthesis code ICARUS[23] to fit both the water-band and broadband modulations under the assumption of no heat redistribution between hemispheres. As shown in Fig. 2, the resulting fit is acceptable and provides a plausible representation of the temperature distribution across the surface of the donor. The average temperature of the dayside is $\langle T_{\mathrm{day}} \rangle = 2{,}401 \pm 10\,\mathrm{K}$ and of the nightside is $\langle T_{\mathrm{night}} \rangle = 2{,}344 \pm 7\,\mathrm{K}$. These numbers are consistent with the mean donor temperature obtained from the template-calibrated spectral-type measurements.

The observed difference between the dayside and nightside fluxes is statistically significant and allows us to estimate the reprocessing efficiency of the donor. This efficiency is usually parameterized via the bolometric (Bond) albedo A_{B}, which is the fraction of the incident irradiation flux that is not reprocessed, but reflected back into space. The irradiating source in the system is the hot white dwarf. We therefore obtained an improved measurement of its temperature, $T_{\mathrm{WD}} = 13{,}200 \pm 200\,\mathrm{K}$, by making use of ultraviolet observations obtained by the GALEX satellite. With this, we find $A_{\mathrm{B}} < 0.54$ at 2σ confidence, as shown in Extended Data Fig. 1. This low value implies a high reprocessing efficiency in the atmosphere of the donor. We can also use the observed difference between dayside and nightside fluxes to estimate the efficiency ϵ with which radiation absorbed locally is redistributed across the donor (see Methods). We find a limit for the redistribution efficiency of $\epsilon < 0.54$ at 2σ confidence. Efficient reprocessing, coupled with poor heat redistribution, has also been found in hot Jupiters in this atmospheric temperature regime[5] and been suggested for low-mass stars in accreting white dwarfs[4]. If irradiation is the main cause of the larger-than-predicted donor radii[19], then efficient heat redistribution is required. Therefore, our finding that the heat transfer from the dayside to the nightside must be modest suggests that irradiation is unlikely to be responsible for inflating the donors. Instead, the dominant effect of irradiation is simply an increase in the local temperature[4].

In most hot Jupiters, the external irradiation overwhelms any internal heat flux. The relatively milder irradiation experienced by the donor in J1433, coupled with the unique prospect of obtaining detailed phase-resolved spectra and line profiles, will permit quantitative tests of irradiated model atmospheres in a regime where these observables are expected to be quite sensitive to irradiation. The system also provides an independent benchmark for theoretical models that predict the albedo and heat redistribution efficiency in irradiated planetary atmospheres[5]. Finally, the donor in J1433 is a rapid rotator (with a projected rotational velocity of $131 \pm 46\,\mathrm{km\,s^{-1}}$), which could have interesting and substantial effects on its atmospheric dynamics[24].

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 17 April 2015; accepted 14 March 2016.

1. Knigge, C. The donor stars of cataclysmic variables. *Mon. Not. R. Astron. Soc.* **373,** 484–502 (2006).
2. Knigge, C., Baraffe, I. & Patterson, J. The evolution of cataclysmic variables as revealed by their donor stars. *Astrophys. J.* Suppl. **194,** 28 (2011).
3. Littlefair, S. P. *et al.* A *J*-band detection of the sub-stellar mass donor in SDSS J1433+1011. *Mon. Not. R. Astron. Soc.* **431,** 2820–2825 (2013).
4. Barman, T. S., Hauschildt, P. H. & Allard, F. Model atmospheres for irradiated stars in precataclysmic variables. *Astrophys. J.* **614,** 338–348 (2004).
5. Perez-Becker, D. & Showman, A. P. Atmospheric heat redistribution on hot Jupiters. *Astrophys. J.* **776,** 134 (2013).
6. Vernet, J. *et al.* X-shooter, the new wide band intermediate resolution spectrograph at the ESO Very Large Telescope. *Astron. Astrophys.* **536,** A105 (2011).
7. Szkody, P. *et al.* Cataclysmic variables from Sloan Digital Sky Survey. VI. The sixth year (2005). *Astron. J.* **134,** 185–194 (2007).
8. Littlefair, S. P. *et al.* A brown dwarf mass donor in an accreting binary. *Science* **314,** 1578–1580 (2006).
9. Savoury, C. D. J. *et al.* Cataclysmic variables below the period gap: mass determinations of 14 eclipsing systems. *Mon. Not. R. Astron. Soc.* **415,** 2025–2041 (2011).
10. Hubeny, I. & Lanz, T. Non-LTE line-blanketed model atmospheres of hot stars. 1: hybrid complete linearization/accelerated lambda iteration method. *Astrophys. J.* **439,** 875–904 (1995).
11. Cushing, M. C., Rayner, J. T. & Vacca, W. D. An infrared spectroscopic sequence of M, L, and T dwarfs. *Astrophys. J.* **623,** 1115–1140 (2005).
12. Burgasser, A. J. *et al.* Not alone: tracing the origins of very-low-mass stars and brown dwarfs through multiplicity studies. In *Protostars and Planets V* (eds Reipurth, B. *et al.*) 427–441 (Univ. Arizona Press, 2007).
13. Zorotovic, M., Schreiber, M. R. & Gänsicke, B. T. Post common envelope binaries from SDSS. XI. The white dwarf mass distributions of CVs and pre-CVs. *Astron. Astrophys.* **536,** A42 (2011).
14. Nordhaus, J. & Spiegel, D. S. On the orbits of low-mass companions to white dwarfs and the fates of the known exoplanets. *Mon. Not. R. Astron. Soc.* **432,** 500–505 (2013).
15. Gänsicke, B. T. *et al.* SDSS unveils a population of intrinsically faint cataclysmic variables at the minimum orbital period. *Mon. Not. R. Astron. Soc.* **397,** 2170–2188 (2009).
16. Politano, M. The formation of cataclysmic variables with brown dwarf secondaries. *Astrophys. J.* **604,** 817–826 (2004).
17. Uthas, H., Knigge, C., Long, K. S., Patterson, J. & Thorstensen, J. The cataclysmic variable SDSS J1507+52: an eclipsing period bouncer in the Galactic halo. *Mon. Not. R. Astron. Soc.* **414,** L85–L89 (2011).
18. Arras, P. & Bildsten, L. Thermal structure and radius evolution of irradiated gas giant planets. *Astrophys. J.* **650,** 394–407 (2006).
19. Littlefair, S. P. *et al.* On the evolutionary status of short-period cataclysmic variables. *Mon. Not. R. Astron. Soc.* **388,** 1582–1594 (2008).
20. Ritter, H., Zhang, Z.-Y. & Kolb, U. Irradiation and mass transfer in low-mass compact binaries. *Astron. Astrophys.* **360,** 969–990 (2000).
21. Ruciński, S. M. The proximity effects in close binary systems. II. The bolometric reflection effect for stars with deep convective envelopes. *Acta Astron.* **19,** 245–255 (1969).
22. Stevenson, K. B. *et al.* Thermal structure of an exoplanet atmosphere from phase-resolved emission spectroscopy. *Science* **346,** 838–841 (2014).
23. Breton, R. P., Rappaport, S. A., van Kerkwijk, M. H. & Carter, J. A. KOI 1224: a fourth bloated hot white dwarf companion found with *Kepler*. *Astrophys. J.* **748,** 115 (2012).
24. Showman, A. P. & Kaspi, Y. Atmospheric dynamics of brown dwarfs and directly imaged giant planets. *Astrophys. J.* **776,** 85 (2013).

Acknowledgements This work is based on observations made with ESO Telescopes at the La Silla Paranal Observatory under programme ID 085.D-0489. J.V.H.S. acknowledges support via studentships from CONACyT (Mexico) and the University of Southampton, as well as research support by the Royal Astronomical Society. R.P.B. has received funding from the European Union Eleventh Framework Programme under grant agreement PIIF-GA-2012-332393. B.T.G. was supported by ERC grant agreement number 320964. P.H.H. acknowledges computing time allocations at the Höchstleistungs Rechenzentrum Nord (HLRN), at the Regionales Rechenzentrum of the University of Hamburg and at the National Energy Research Supercomputer Center (NERSC), which is supported by the Office of Science of the US Department of Energy under contract number DE-AC03-76SF00098.

Author Contributions C.K., S.P.L., V.S.D., B.T.G., T.R.M., M.L.P. and J.S. proposed and planned the observations. All data analysis was done by J.V.H.S. with substantial feedback from C.K., R.P.B. and S.P.L. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.V.H.S. (j.v.hernandez@soton.ac.uk).

© 2016 Macmillan Publishers Limited. All rights reserved

METHODS

Mass determination. We performed a radial velocity analysis on the 1.243-μm and 1.252-μm K I absorption lines formed in the atmosphere of the donor. The trailed spectrum of these features is shown as the inset in Fig. 1. We performed a simultaneous fit to both lines and found the best values for the semi-amplitude $K_2 = 499 \pm 15\,\mathrm{km\,s^{-1}}$, the systemic velocity $\gamma = -42 \pm 8\,\mathrm{km\,s^{-1}}$ and the rotational broadening $v\sin i = 131 \pm 46\,\mathrm{km\,s^{-1}}$. When combined with the observed value for the radial velocity semi-amplitude of the white dwarf, $K_1 = 34 \pm 4\,\mathrm{km\,s^{-1}}$ (ref. 25), these best-fit values allow us to derive a purely dynamical estimate of the mass ratio of the system: $q = M_2/M_1 = K_1/K_2 = 0.068 \pm 0.008$. For this mass ratio, the occurrence of white-dwarf eclipses alone sets a lower limit on the inclination of $i > 80°$ (ref. 26). We can then use the mass function of the secondary, $f(M_2) = P_{\mathrm{orb}}K_1^3/(2\pi G) = M_1\sin^3 i/(1+q)^2$ where G is the gravitational constant, to set a purely dynamical 2σ upper limit on the mass of the donor star, $M_2 < 0.071 M_\odot$. We can also use the mass function to obtain a model-independent estimate of the donor mass. In any semi-detached compact binary system, there is a unique, single-valued family of i–q pairs that produce a compact object eclipse with a given width[26]. In the case of J1433, the width of the white-dwarf eclipse can be measured directly from published optical light curves[19]. We can then combine the resulting i–q constraint with our spectroscopic mass ratio and the mass function to obtain a robust estimate of the donor mass that depends only on Kepler's third law and the well-understood geometry of Roche-lobe-filling objects. We find $M_2 = (0.055 \pm 0.008) M_\odot$, well below the hydrogen-burning limit.

Water-band ratios and broadband fluxes. To provide the cleanest measurements of the donor at any given phase, we subtracted contributions from both the white dwarf and the accretion disk. The flux observed from the white dwarf was assumed to be constant in phase, except for phase bins affected by an eclipse. In these, we reduced the flux by the fraction of the bin width during which the white dwarf was occulted. The contribution from the accretion disk was modelled by fitting a power law to the optical (3,200–5,400 Å) wave band (while masking the emission lines) and extrapolating to the near-infrared. Owing to the low signal-to-noise ratio of the individual spectra, the telluric removal created artefacts in some regions of the spectra. We identified these regions and masked them before shifting each spectrum to the rest frame of the donor. We binned the data into eight orbital phase bins and measured a slightly modified water-band index, defined as $\langle F_{\lambda=1.318\,\mu\mathrm{m}}\rangle/\langle F_{\lambda=1.340\,\mu\mathrm{m}}\rangle$, where the fluxes F_λ are averages over a 5-Å window. This index was then calibrated against brown-dwarf spectral-type templates[11]. Errors were estimated via 500 bootstrap copies of the median flux inside every orbital bin for all the spectra. The broadband flux modulation was estimated by integrating the donor flux over the J band (after masking all emission lines) and converting to AB magnitudes.

Dayside and nightside temperatures, albedo and intrinsic luminosity. To model the orbital-phase-dependent water-band ratios and broadband flux, we constructed a grid of irradiated donor models with ICARUS[23]. Each of these models is characterized by two parameters: the irradiating luminosity absorbed by the donor and the intrinsic luminosity it would produce in the absence of irradiation. For each parameter pair, we use ICARUS to calculate the resulting temperature distribution across the donor and to predict the observed spectrum at each orbital phase. These calculations assume that all of the irradiating flux absorbed at a given point on the surface of the donor is re-emitted locally and neglect any effect of the fast rotation of the donor on the temperature distribution (which could affect the atmospheric dynamics[24]). However, the distorted shape of the Roche-lobe-filling donor, as well as limb- and gravity-darkening, are fully taken into account in these calculations. The absorbed irradiating and intrinsic luminosities are not directly observable, but they determine the dayside and nightside temperatures, respectively, which are observable. Therefore, we present our results in terms of these temperatures, which we calculate by summing over all dayside surface elements in the model:

$$\langle T_{\mathrm{day}}\rangle = \left(\frac{1}{\sigma}\frac{\sum_{j,\mathrm{day}} F_{\mathrm{tot},j}}{\sum_{j,\mathrm{day}} A_j}\right)^{1/4}$$

and all nightside elements:

$$\langle T_{\mathrm{night}}\rangle = \left(\frac{1}{\sigma}\frac{\sum_{j,\mathrm{night}} F_{\mathrm{int},j}}{\sum_{j,\mathrm{night}} A_j}\right)^{1/4}$$

Here, $F_{\mathrm{int},j} = A_j\sigma T_{\mathrm{int},j}^4$ is the intrinsic flux, where A_j is the area of surface element j and σ is the Stefan–Boltzmann constant. As a consequence of the distorted shape of the donor, $T_{\mathrm{int},j}$—the temperature of each surface element in the absence of irradiation—is not constant. Similarly, the total flux that heats each surface element on the dayside, $F_{\mathrm{tot},j} = F_{\mathrm{int},j} + (1 - A_{\mathrm{B}})F_{\mathrm{irr},j}$, where A_{B} is the bolometric (Bond) albedo of the donor. $F_{\mathrm{irr},j}$ is the irradiating flux seen by this element:

$$F_{\mathrm{irr},j} = \frac{4\pi R_{\mathrm{WD}}^2 \sigma T_{\mathrm{WD}}^4}{4\pi d_j^2} A_{\mathrm{proj},j}$$

in which d_j is the distance of surface element j to the white dwarf, T_{WD} is the temperature of the white dwarf, R_{WD} is the radius of the white dwarf and $A_{\mathrm{proj},j}$ is the projected surface intercepting the irradiating flux. Because we have good estimates of the temperature of the white dwarf and its binary parameters, $F_{\mathrm{irr},j}$ is well constrained observationally. Therefore, the absorbed irradiating flux and the dayside temperature can equivalently be thought of as measures of the albedo. We determine optimal values of $\langle T_{\mathrm{day}}\rangle$, $\langle T_{\mathrm{night}}\rangle$ and A_{B} by fitting the observed phase-dependent water-band and broadband fluxes using a Markov chain Monte Carlo procedure. The results are shown in the Extended Data Fig. 1. We find $\langle T_{\mathrm{night}}\rangle = 2{,}344^{+11}_{-10}\,\mathrm{K}$ and $\langle T_{\mathrm{day}}\rangle = 2{,}401^{+7}_{-7}\,\mathrm{K}$; the errors represent the 1σ confidence levels. The corresponding constraints on the albedo and the intrinsic donor luminosity are $A_{\mathrm{B}} < 0.54$ (at 2σ confidence) and $L_{\mathrm{int}} = (3.1 \pm 0.1) \times 10^{-4} L_\odot$ (at 1σ confidence), respectively. In all of these models, we adopt a white-dwarf temperature of $T_{\mathrm{WD}} = 13{,}200 \pm 200\,\mathrm{K}$ on the basis of our own analysis of ultraviolet data obtained by GALEX, an inclination of $i = 84.36°$ and a distance to the system of $d = 226$ pc, and adopt an orbital period of $P_{\mathrm{orb}} = 78.106657$ min on the basis of previous eclipse modelling estimates[9]. We also adopt a fixed gravity-darkening coefficient of $\beta = 0.08$ (ref. 27).

Heat redistribution across the terminator. Our geometric reprocessing model assumes that none of the irradiating flux is redistributed from the dayside to the nightside. This model fits the data acceptably, suggesting that redistribution is inefficient in the atmosphere of the donor. To estimate this efficiency quantitatively, we need to adopt a specific model for redistribution. Here, we assume that a constant fraction $0 \le \epsilon \le 1$ of the irradiating flux that is absorbed at any point on the dayside of the donor is redistributed evenly across the entire surface. Similar models are commonly used in studies of irradiated exoplanets[28]. In the context of this simple model, the difference between the fluxes emerging from the dayside and the nightside is:

$$\sigma\langle T_{\mathrm{day}}\rangle^4 - \sigma\langle T_{\mathrm{night}}\rangle^4 = (\sigma\langle T_{\mathrm{day}}\rangle^4 - \sigma\langle T_{\mathrm{night}}\rangle^4)_{\epsilon=0} - \epsilon(1 - A_{\mathrm{B}})\frac{\sum_{j,\mathrm{day}} F_{\mathrm{irr},j}A_j}{\sum_{j,\mathrm{day}} A_j}$$

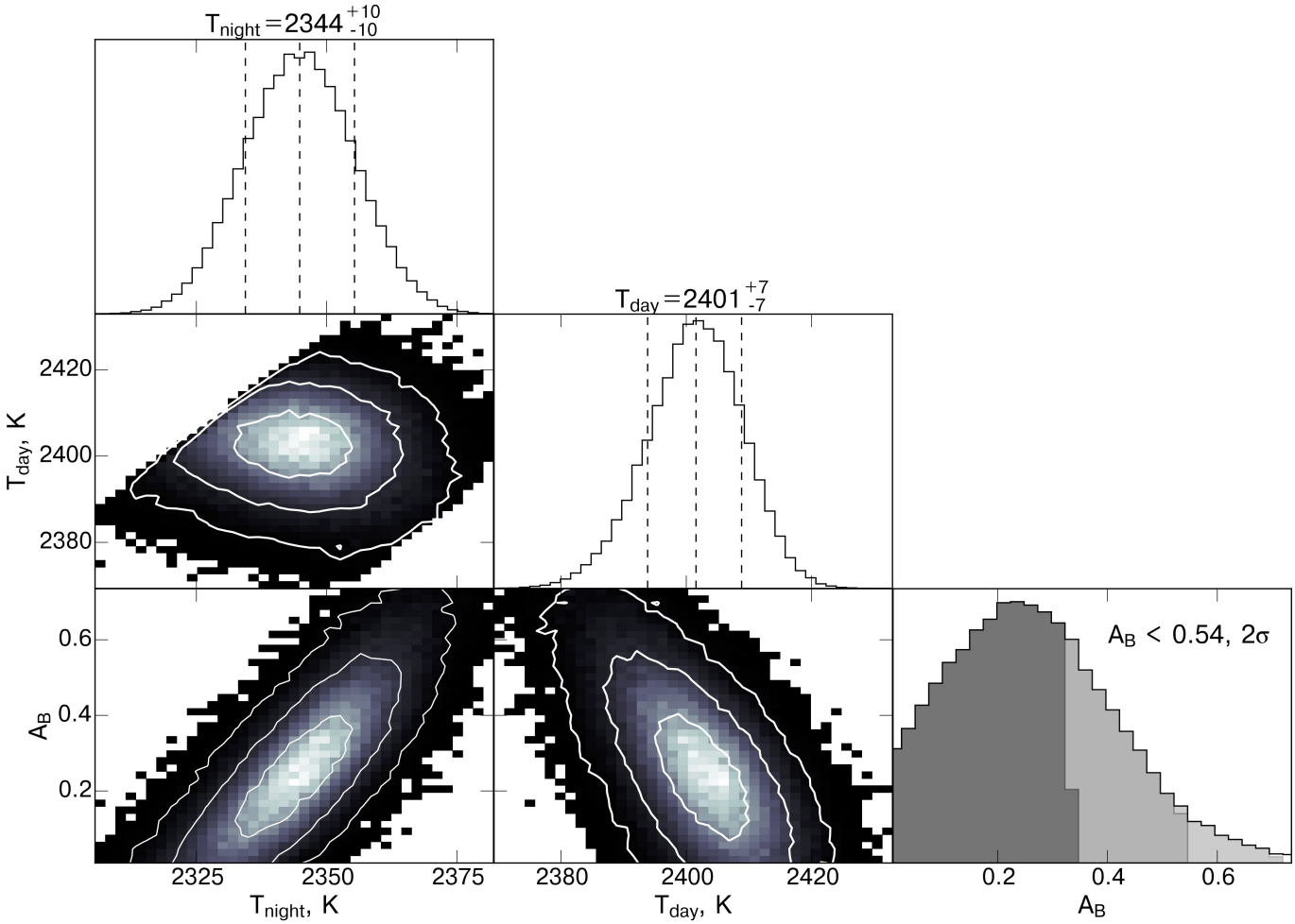
in which $(\sigma\langle T_{\mathrm{day}}\rangle^4 - \sigma\langle T_{\mathrm{night}}\rangle^4)_{\epsilon=0}$ is the flux difference of the equivalent model without any redistribution. The modelling described in the previous section provides us with a set of self-consistent flux differences for models with $\epsilon = 0$. We can therefore transform these to the flux differences predicted for otherwise equivalent models with $\epsilon > 0$. Because our earlier modelling also provides us with estimates of the observed dayside and nightside temperatures, we can estimate ϵ by comparing the observationally inferred flux difference to that predicted by our transformed models with $\epsilon > 0$. The results are shown in Extended Data Fig. 2. We find that the efficiency of redistribution is indeed low, $0.1 < \epsilon < 0.4$, with higher values corresponding to lower albedos. The 2σ upper limit on the redistribution efficiency (for $A_{\mathrm{B}} = 0$) is $\epsilon < 0.54$.

Sample size. No statistical methods were used to predetermine sample size.

Code availability. The code used to generate the model, as well as the atmosphere templates, ICARUS, is available at http://icarus.renebreton.org.

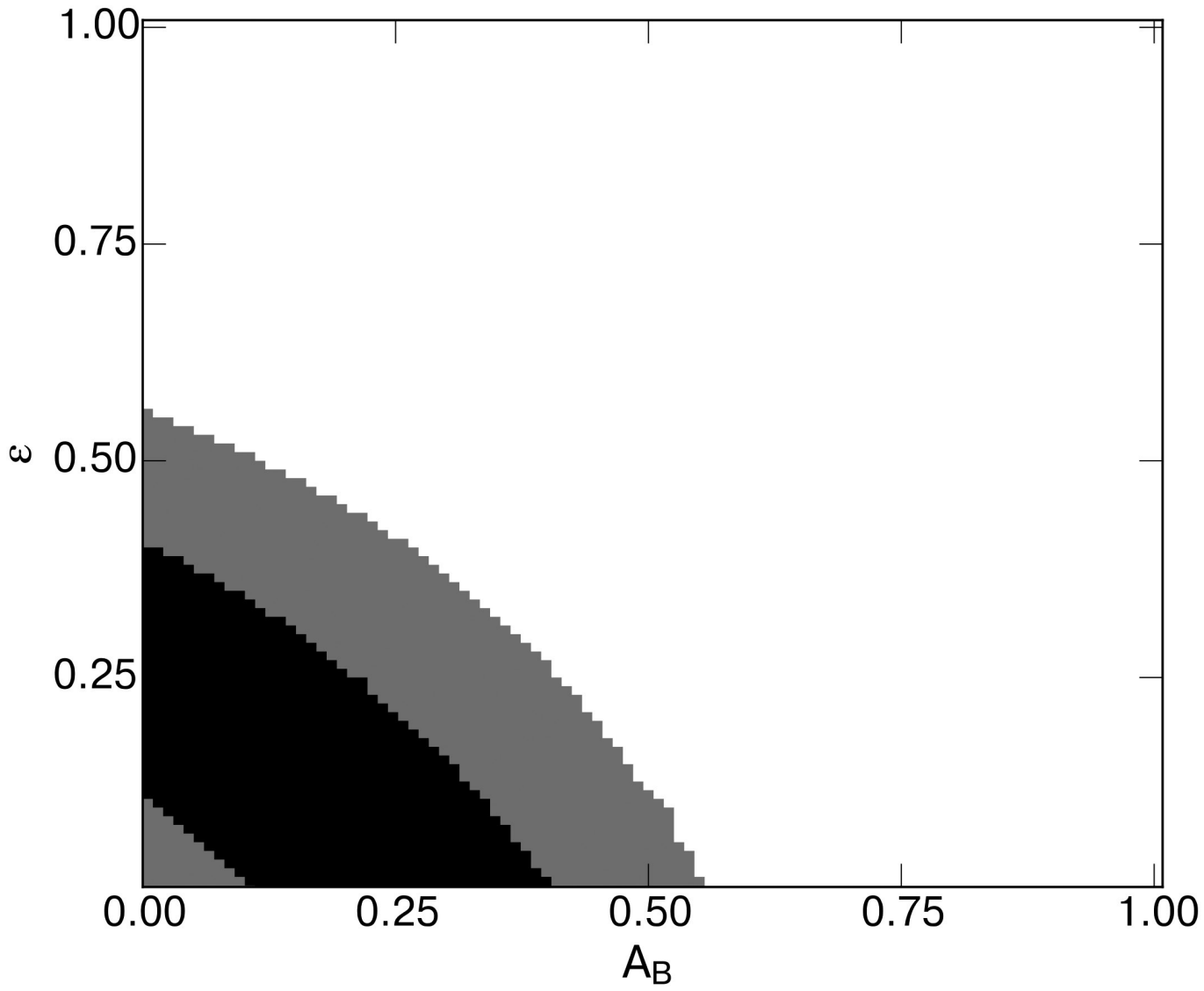
25. Tulloch, S. M., Rodríguez-Gil, P. & Dhillon, V. S. Radial velocity study of the post-period minimum cataclysmic variable SDSS J143317.78+101123.3 with an electron-multiplying CCD. *Mon. Not. R. Astron. Soc.* **397,** L82–L86 (2009).
26. Chanan, G. A., Middleditch, J. & Nelson, J. E. The geometry of the eclipse of a pointlike star by a Roche-lobe-filling companion. *Astrophys. J.* **208,** 512–517 (1976).
27. Lucy, L. B. Gravity-darkening for stars with convective envelopes. *Z. Astrophys.* **65,** 89–92 (1967).
28. Cowan, N. B. & Agol, E. The statistics of albedo and heat recirculation on hot exoplanets. *Astrophys. J.* **729,** 54 (2011).

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 1 | Posterior probability distributions for the irradiation model parameters. Colour scale contours show the joint probability for every combination of parameters. Contours represent the 1σ, 2σ and 3σ levels. Marginal posterior distributions are shown as histograms with the median and 1σ marked as dashed lines. The A_B distribution is quoted as a truncated distribution with a 2σ upper limit.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 2 | Redistribution efficiency limits for the irradiated substellar donor. Allowed family solutions of redistribution efficiencies ε as a function of Bond albedo A_B are shown. The black and grey contours represent the 1σ and 2σ confidence levels, respectively.

© 2016 Macmillan Publishers Limited. All rights reserved

LETTER

doi:10.1038/nature17633

Self-assembly of coherently dynamic, auxetic, two-dimensional protein crystals

Yuta Suzuki[1], Giovanni Cardone[1], David Restrepo[2], Pablo D. Zavattieri[2], Timothy S. Baker[1,3] & F. Akif Tezcan[1]

Two-dimensional (2D) crystalline materials possess unique structural, mechanical and electronic properties[1,2] that make them highly attractive in many applications[3–5]. Although there have been advances in preparing 2D materials that consist of one or a few atomic or molecular layers[6,7], bottom-up assembly of 2D crystalline materials remains a challenge and an active area of development[8–10]. More challenging is the design of dynamic 2D lattices that can undergo large-scale motions without loss of crystallinity. Dynamic behaviour in porous three-dimensional (3D) crystalline solids has been exploited for stimuli-responsive functions and adaptive behaviour[11–13]. As in such 3D materials, integrating flexibility and adaptiveness into crystalline 2D lattices would greatly broaden the functional scope of 2D materials. Here we report the self-assembly of unsupported, 2D protein lattices with precise spatial arrangements and patterns using a readily accessible design strategy. Three single- or double-point mutants of the C_4-symmetric protein RhuA were designed to assemble via different modes of intermolecular interactions (single-disulfide, double-disulfide and metal-coordination) into crystalline 2D arrays. Owing to the flexibility of the single-disulfide interactions, the lattices of one of the variants (^{C98}RhuA) are essentially defect-free and undergo substantial, but fully correlated, changes in molecular arrangement, yielding coherently dynamic 2D molecular lattices. ^{C98}RhuA lattices display a Poisson's ratio of −1—the lowest thermodynamically possible value for an isotropic material—making them auxetic.

Proteins are attractive building blocks for 2D materials because of their structural and chemical diversity, and inherent functions. Examples for natural, protein-based 2D materials include bacterial S-layer proteins and purple-membrane assemblies, which form crystalline arrays in association with cell walls and membranes, respectively, and have been used in diverse technological applications[14,15]. On the synthetic front, methods for 2D protein crystallization have been developed for the structural characterization of membrane proteins[16,17] or functional applications[18], and generally rely on lipid layers as supports. Recently, 2D or 3D supramolecular protein arrays have been designed through the symmetric polymerization of protein building blocks via computationally designed protein–protein interactions or fusion of protein components[19–22]. However, these approaches are engineering-intensive and highly dependent on the accuracy of the design, and the integration of dynamic or adaptive behaviour has yet to be explored.

To address these issues, we used a simple chemical-bonding strategy to control protein self-assembly. We reasoned that both cysteine (Cys)-mediated disulfide bonds and metal-coordination interactions between protein building blocks could produce crystalline and dynamic arrays with minimal design, because these bonds are: (1) strong but reversible (to minimize the surface area to be designed and ensure self-healing); (2) short, yet sufficiently flexible (to simultaneously afford crystallinity and adaptiveness); (3) chemically tunable (to exert external control over self-assembly and enable stimuli-responsiveness); and (4) easily designed and engineered.

The most straightforward route to obtaining 2D lattices is the tessellation of C_3-, C_4- or C_6-symmetric building blocks through appropriately positioned C_2-symmetric linkages such as disulfide bonds or many metal-coordination interactions. As a model building block, we chose L-rhamnulose-1-phosphate aldolase (RhuA), a C_4-symmetric homotetramer (dimensions, 7 nm × 7 nm × 5 nm) that was previously used as a synthon for supramolecular assemblies[23] (Fig. 1a). An inspection of RhuA indicated protrusions in the four corner positions as ideal locations to incorporate one Cys or two histidine (His) residues for disulfide- or metal-directed self-assembly, respectively. We generated two variants, ^{C98}RhuA and $^{H63/H98}$RhuA, with four conditionally self-associating corners for forming square lattices (Fig. 1a, b, top and middle rows). The relative positions of residues 63 and 98 were deemed to be conducive to forming a bis-His metal-coordination motif to afford a tetrahedral or square planar coordination geometry for metal-mediated RhuA pairing interactions. Additionally, we noted that RhuA could be converted into a stable D_4-symmetric octamer via a single point mutation (A88F)[24]. Thus, we prepared a third variant, $^{F88/C98}$RhuA, which presents eight symmetry-related cysteines at roughly 45° angles in a 2D projection (Fig. 1a, bottom row). We envisioned that the C_4-symmetric ^{C98}RhuA and $^{H63/H98}$RhuA variants could yield square lattices with two distinct patterns in terms of the orientation of the building blocks with respect to the 2D plane (Fig. 1b). On the other hand, the D_4 symmetry of $^{F88/C98}$RhuA would dictate a 2D lattice with equivalent faces (Fig. 1b). These RhuA building blocks would also provide three distinct modes of inter-building-block interactions, whose effects on self-assembly we hoped to investigate: single-disulfide, double-disulfide and bis-His-anchored metal-coordination.

For the oxidative self-assembly of ^{C98}RhuA and $^{F88/C98}$RhuA, we tested various strategies including air oxidation, redox buffer systems containing reduced and oxidized glutathione (GSH and GSSG) or low concentrations (≤10 mM) of the reductant β-mercaptoethanol (βME), which slowly decomposes in aqueous solution and results in gradually more oxidizing conditions. In these experiments, solutions of purified ^{C98}RhuA and $^{F88/C98}$RhuA were first rapidly exchanged via repeated centrifugal filtration into solutions that varied in terms of their pH (6–8.5), buffering species (sodium phosphate, CHES, MES, MOPS, Tris; at 5–20 mM), the compositions of the redox buffering system (1:19 or 19:1 GSH:GSSG; at 1 mM total concentration) or the concentrations of βME (0–10 mM). After adjustment to the desired final protein concentration (25–175 μM), these solutions were monitored for the formation of self-assembled structures by visual inspection for emergence of cloudiness and by transmission electron microscopy (TEM). Likewise, for metal-directed assembly, purified $^{H63/H98}$RhuA was exchanged into solutions with varying pH and buffer ions (to modulate metal binding kinetics and thermodynamics), with the exception that these solutions did not contain any reductants.

[1]Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, California 92093, USA. [2]School of Civil Engineering, Purdue University, West Lafayette, Indiana 47907-2051, USA. [3]Division of Biological Sciences, University of California, San Diego, La Jolla, California 92093, USA.

© 2016 Macmillan Publishers Limited. All rights reserved

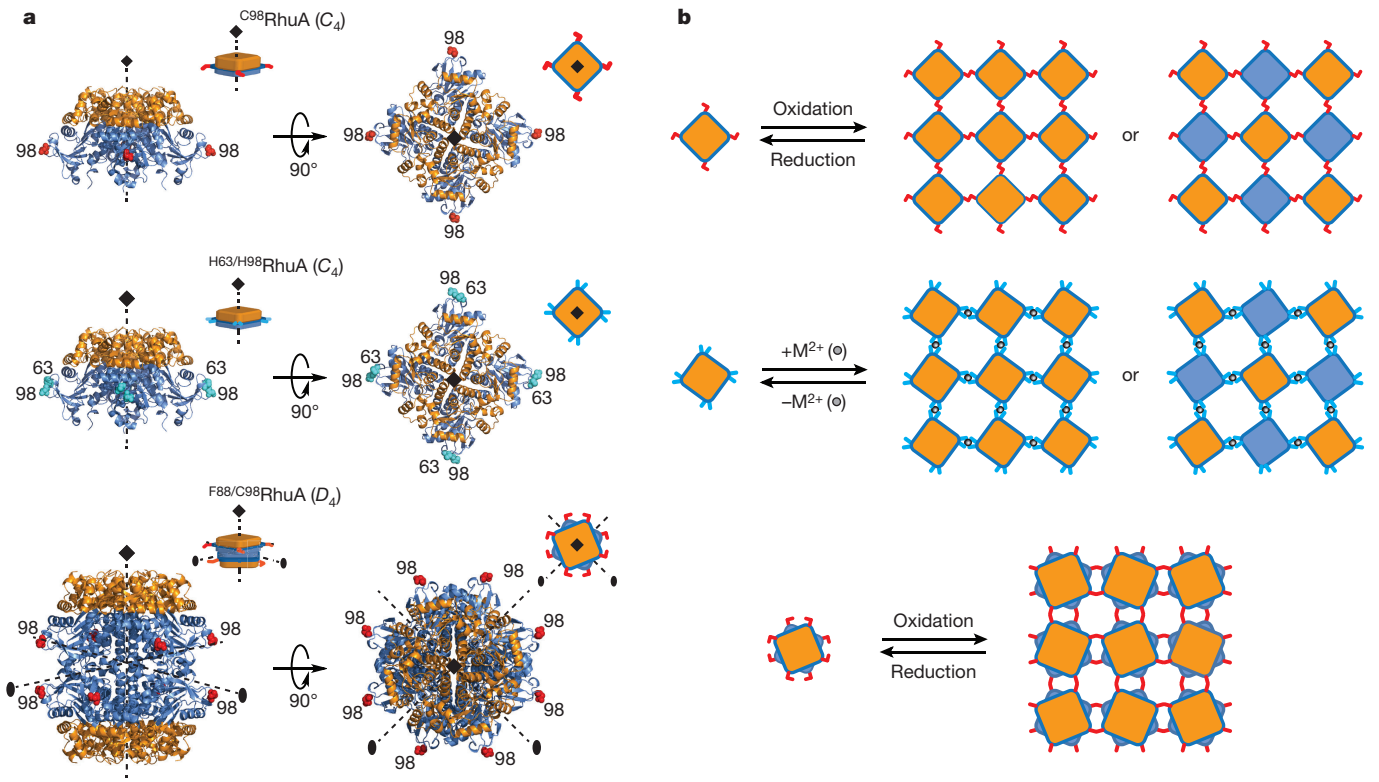


Figure 1 | RhuA constructs and their disulfide- and metal-mediated self-assembly modes. a, Schematic representations of the ^{C98}RhuA, $^{H63/H98}$RhuA and $^{F88/C98}$RhuA structures. The top and bottom halves of RhuA are coloured orange and blue, respectively, to highlight the relative molecular orientations in **b**. Cys and His residues inserted into positions 98 and 63 are shown in red and cyan, respectively. **b**, Expected 2D molecular arrangements of ^{C98}RhuA, $^{H63/H98}$RhuA and $^{F88/C98}$RhuA lattices. M^{2+} refers to Zn^{2+} or Cu^{2+} ions.

Self-assembly was initiated by addition of 4–40 molar equivalents of Zn^{2+} and Cu^{2+} (1–10 equiv. per bis-His motif), both of which can accommodate the desired four-coordinate geometries to link $^{H63/H98}$RhuA building blocks at their corners.

We found that the assembly of all variants into 2D arrays was robust and occurred under a wide range of conditions. As expected, crystalline self-assembly was favoured by conditions that promoted slow, controlled oxidation or slow metal-binding kinetics. For example, uncontrolled air oxidation of ^{C98}RhuA solutions or addition of large molar excess of Zn^{2+} and Cu^{2+} to $^{H63/H98}$RhuA samples resulted largely in amorphous aggregates in addition to some crystalline domains (Extended Data Fig. 1). The following solution conditions yielded 2D assemblies that were optimized in terms of size, crystallinity and yield: ^{C98}RhuA and $^{F88/C98}$RhuA ($\geq$125 μM protein, 10 mM βME, pH 7.5, 10 mM Tris); $^{H63/H98}$RhuA (25 μM protein, 200 μM $ZnCl_2$, pH 7, 20 mM MOPS). Under these conditions, ^{C98}RhuA reproducibly assembled into straight-edged, single- or few-layered 2D crystals that grew to several micrometres in size over several days (Fig. 2a). Negative-stain TEM, cryo-TEM, scanning electron microscopy (SEM) and atomic force microscopy (AFM) measurements revealed that the ^{C98}RhuA crystals were highly ordered and possessed uniform square or rectangular shapes with molecularly sharp boundaries (Fig. 2a, Extended Data Figs 2 and 3); these features have not been previously observed in designed 2D protein crystals[19,21]. 2D crystals of $^{H63/H98}$RhuA (Fig. 2b, Extended Data Fig. 4) and $^{F88/C98}$RhuA (Fig. 2c, Extended Data Fig. 5) were several hundred nanometres to one micrometre in size, but typically displayed irregular morphologies. $^{H63/H98}$RhuA lattices were monocrystalline, but tended to grow in three dimensions over time, whereas $^{F88/C98}$RhuA crystals consisted of polycrystalline domains.

As monitored by TEM and dynamic light scattering (DLS), self-assembly of all variants was reversible upon addition of high concentrations of either βME (>10 mM) in the case of ^{C98}RhuA and $^{F88/C98}$RhuA crystals or ethylene-diaminetetraacetic acid (EDTA; >10 mM) in the case of $^{H63/H98}$RhuA (Extended Data Fig. 6). The variant ^{D98}RhuA, which bears the native D98 residue and is devoid of corner Cys residues or the bis-His motifs, was not competent to self-assemble upon oxidation or metal addition (Extended Data Fig. 7a). These observations confirm the proposed modes of disulfide- or metal-mediated self-assembly. Two additional variants (^{C133}RhuA and ^{C266}RhuA) were prepared with surface cysteines located on the sides of the tetramers (Extended Data Figs 7b, c). Theoretically, these variants could be envisioned to self-associate into close-packed 2D square lattices through disulfide linkages. Instead, they both formed non-crystalline aggregates upon oxidation, which we ascribe to the steric crowding around the C133 and C266 positions that precludes the orientation of RhuA molecules for planar self-assembly. These experiments suggest that the placement of Cys residues at sterically isolated, corner locations is preferable for the formation of 2D crystalline arrays.

Exhaustive TEM analyses involving many crystals of each variant indicated $p42_12$ plane group symmetry for ^{C98}RhuA and $^{F88/C98}$RhuA and $p4$ symmetry for $^{H63/H98}$RhuA crystals (Fig. 2, Extended Data Fig. 8a, b and Supplementary Table 4). These assignments were corroborated by electron diffraction data for ^{C98}RhuA and $^{H63/H98}$RhuA crystals (Extended Data Fig. 8c). Experimental 2D electron microscopy projection maps agree very well with those calculated from structural models (Extended Data Fig. 9a) and confirm the expected protein–protein interaction geometries: corner-attached ^{C98}RhuA and $^{H63/H98}$RhuA and side-attached $^{F88/C98}$RhuA (Fig. 2, columns (iv) and (v)). The plane group symmetries suggest that ^{C98}RhuA tetramers adopt a checkerboard pattern ($p42_12$ symmetry), whereas $^{H63/H98}$RhuA tetramers self-assemble in the same orientation with respect to the 2D plane ($p4$ symmetry). Together with the fact that $^{F88/C98}$RhuA possesses two-fold symmetry parallel to the 2D plane, we thus obtain three distinct 2D molecular patterns with minimal

© 2016 Macmillan Publishers Limited. All rights reserved

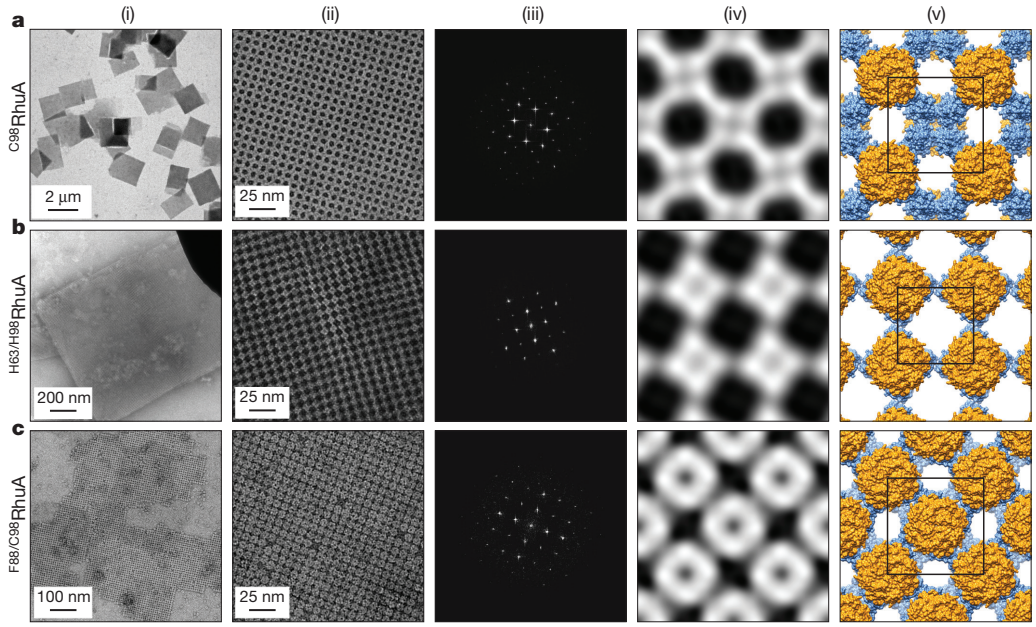


Figure 2 | TEM characterization of the 2D crystals of RhuA variants. **a**, ^{C98}RhuA; **b**, $^{H63/H98}$RhuA (Zn^{2+} coordination); and **c**, $^{F88/C98}$RhuA. (i) Low-magnification views, (ii) high-magnification views, (iii) Fourier transforms of column (ii), (iv) reconstructed 2D images, and (v) structural models based on the 2D reconstructions in column (iv) (colouring as in Fig. 1). The high-resolution limits in column (iii) are approximately 14 Å. Unit cells are shown as black squares in column (v) (with corners placed at four-fold symmetry axes); unit cell constants ($\alpha = \beta, \gamma$): ^{C98}RhuA (114 Å, 90°), $^{H63/H98}$RhuA (91 Å, 90°) and $^{F88/C98}$RhuA (115 Å, 90°).

alteration of a single protein building block (Extended Data Fig. 9b). ^{C98}RhuA and $^{H63/H98}$RhuA molecules self-assemble in distinct orientations in the 2D plane, suggesting that the energetic bias stemming from disulfide or metal bonding configurations and surrounding protein–protein interactions must be sufficiently large to favour one orientation over others to yield long-range order.

Superior quality and larger sizes of ^{C98}RhuA crystals compared to $^{F88/C98}$RhuA and $^{H63/H98}$RhuA lattices are readily explained by interactions that direct the self-assembly of the variants. $^{F88/C98}$RhuA presents a roughly circular distribution of eight cysteines (Fig. 1a, bottom row). This arrangement renders the desired double-disulfide-mediated, side-to-side self-assembly mode non-unique, and permits alternate attachment geometries between $^{F88/C98}$RhuA molecules (Extended Data Fig. 5). Moreover, the increased Cys-valency of $^{F88/C98}$RhuA leads to considerably stronger interactions and therefore 'stiffer' lattices: the dissolution of $^{F88/C98}$RhuA polycrystallites requires substantially higher amounts of βME and longer incubation periods compared to ^{C98}RhuA lattices (Extended Data Fig. 6). Consequently, $^{F88/C98}$RhuA lattices display vacancies, as well as both high- (>30°) and low-angle (<10°) grain boundaries (Extended Data Fig. 5). $^{H63/H98}$RhuA crystals possess fewer such defects, owing to the reversibility of Zn^{2+} coordination interactions. Yet, each $^{H63/H98}$RhuA building block also contains numerous surface residues that can weakly coordinate Zn^{2+} ions, promoting crystal growth in the third dimension (Extended Data Fig. 4a).

In contrast to $^{F88/C98}$RhuA and $^{H63/H98}$RhuA, self-assembly of ^{C98}RhuA is both chemically and orientationally specific. Moreover, the reversibility and inherent flexibility of single Cys-Cys linkages—containing five rotatable bonds—probably allows for the correction of any defects such as vacancies and grain boundaries, which would be difficult to accomplish in a rigid lattice composed of strongly interacting building blocks such as $^{F88/C98}$RhuA. Indeed, as TEM images of mono-layered ^{C98}RhuA crystals illustrate, the outcomes are (a) macroscopic crystal morphologies that reflect the molecular symmetry of the building blocks, (b) molecularly sharp crystal boundaries, and (c) lattices with extremely low defect frequencies. In the hundreds of monocrystalline ^{C98}RhuA lattices with surface areas of >1 μm^2 that we examined closely, we rarely found a lattice defect, despite the fact that these crystals grow in 3D space in an unsupported fashion. Even in a rare instance such as that shown in Extended Data Fig. 9d, the defect frequency was one missing ^{C98}RhuA molecule within a lattice grid of about 9,000 molecules (about 0.6 μm^2).

A more striking consequence of the disulfide bond flexibility is the coherent dynamic behaviour of ^{C98}RhuA crystals. Although our initial sample preparations of ^{C98}RhuA crystals yielded predominantly open lattices with large pores (Fig. 2a), we noticed that these crystal suspensions developed a dense sediment over a period of 1–3 days at 4 °C (Extended Data Fig. 10a). TEM analysis of these sedimented crystals indicated a close-packed lattice arrangement (Fig. 3a, b, right panels). Upon resuspension of the sedimented crystals by repeated gentle mixing with a pipette and subsequent TEM imaging of the resulting samples, we captured a total of at least seven types of 2D ^{C98}RhuA crystals in distinct conformational states (I–VII) (Fig. 3a, Extended Data Fig. 10b). These conformational states were categorized by computational image analysis according to the roundness indices of the lattice pores, ranging from ≥0.85 for state I to ≤0.35 for state VII (see Methods). As evidenced by the retention of $p42_12$ symmetry and nearly equal unit-cell dimensions, these seven conformational states are clearly interconnected and implicate a continuous lattice motion between fully open and fully closed states (Supplementary Video 1). These large amplitude motions of the ^{C98}RhuA lattices are afforded by a remarkable extent of hinging about the flexible disulfide linkages and their placement at corner locations of RhuA molecules. The transition from the open to the closed state is accompanied by the compression of the inter-^{C98}RhuA hinge angle (α) from >80° to 17°, a decrease of the pore size from approximately 4.4 nm to 1.0 nm (for the passage of a spherical object) and an increase in the relative protein/hole surface density of 170% (Fig. 3b, Table 1).

Conformational dynamics of ^{C98}RhuA crystals are fully coherent: in each crystal examined, only one type of conformational state (I–VII) was observed throughout the lattice (Fig. 3b). This observation implies that any mechanical deformation of a ^{C98}RhuA crystal is cooperatively propagated along the 2D plane, enabled by both the flexibility and short linker-length of single-disulfide linkages: longer, more flexible linkages would preclude coherent dynamics, whereas inflexible linkages would lead to non-adaptive lattices. Indeed, $^{F88/C98}$RhuA and $^{H63/H98}$RhuA lattices do not display any apparent dynamic behaviour

© 2016 Macmillan Publishers Limited. All rights reserved

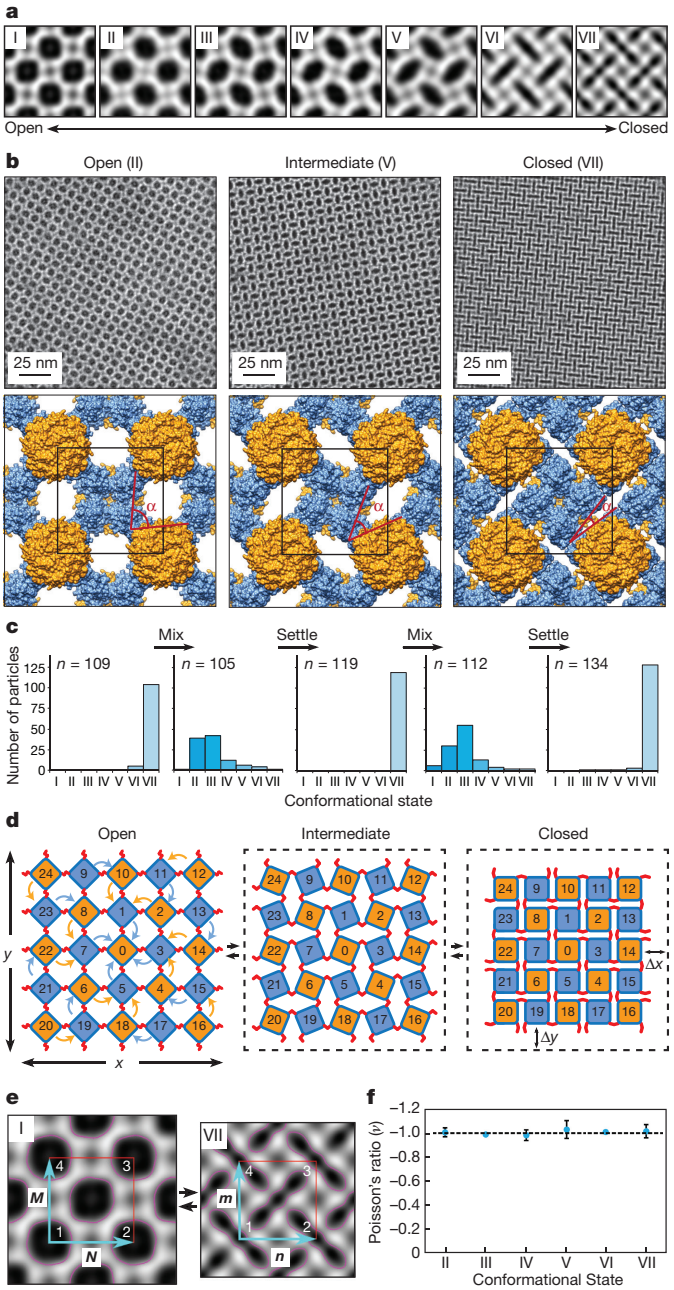


Figure 3 | Dynamic and auxetic nature of 2D ^{C98}RhuA crystals. **a**, Reconstructed 2D images of seven distinct conformational states (I–VII) of 2D ^{C98}RhuA crystals. **b**, High-magnification views and derived structural models of conformations II, V and VII. Unit cells and hinge angles (α) between ^{C98}RhuA units are highlighted in black and red, respectively (orange and blue colouring as in Fig. 1). **c**, Population distributions of ^{C98}RhuA crystals in different conformational states during repeated resuspension/sedimentation cycles. *n* refers to the total number of individual lattices analysed in each panel. **d**, Schematic representation of the rotating, rigid-square model that describes the 2D ^{C98}RhuA lattice dynamics (colouring as in Fig. 1; the building blocks are numbered for clarity). Δx and Δy denote changes in the transverse- and longitudinal-dimension lengths upon lattice opening and closing. **e**, Digital image-correlation analysis of reconstructed TEM images of ^{C98}RhuA crystals for determining auxetic behaviour. Representative volume elements (RVEs) in the lattices of the two extreme states (I and VII) are indicated with red squares, with the vertices of the RVEs numbered 1–4 , and the vectors ***M***, ***N*** (in state I) and ***m***, ***n*** (in state VII) used to calculate local engineering strains are shown with cyan arrows. The edges of the lattice pores, used for determining the positions of the RVE vertices, are shown as purple lines. **f**, Calculated Poisson's ratios (ν) of lattice conformational states with respect to state I. The error bars correspond to the uncertainties in pixel selection during digital image processing (see Methods for details).

Table 1 | Structural parameters of ^{C98}RhuA crystals in states II, V and VII

	State II	State V	State VII
Unit cell ($a=b,\gamma$)	114.4 Å, 90°	110 Å, 90°	107.8 Å, 90°
Plane group symmetry	$p42_12$	$p42_12$	$p42_12$
Intertetramer hinge angle, α	79°	49°	17°
Protein surface area per unit cell, $A_{protein}$	82.7 nm^2	86.2 nm^2	88.0 nm^2
Pore surface area per unit cell, A_{pore}	44.7 nm^2	34.7 nm^2	28.2 nm^2
Relative protein/pore density, $A_{protein}/A_{pore}$	1.9	2.5	3.1

because their double-disulfide- or metal-mediated modes of assembly do not allow rotation of the neighbouring protein molecules with respect to one another.

To establish that the conformational dynamics of ^{C98}RhuA crystals are reversible, we subjected them to repeated sedimentation/resuspension cycles within the solutions in which they self-assembled at 4 °C. In each cycle, we obtained TEM images of >100 individual crystals from the same container, collected from either the sediments that formed overnight or the suspensions immediately after mixing the sediments by repeated pipetting. As shown in Fig. 3c, the conversion between the closed states (VI and VII) and open or intermediate states (I–V) is completely reversible and absolute. Sedimented samples do not contain any open or intermediate states and resuspended samples do not contain any closed states, providing unambiguous evidence that ^{C98}RhuA lattices are reversibly dynamic. On the basis of the observed distribution profiles, the opening of the lattices by mechanical agitation and their distribution among conformational states I–V appear to be immediate (at least within the ~5-min timescale of TEM sample preparation), whereas their full closure takes several hours. We posit that the energetic barriers between different ^{C98}RhuA lattice conformations (due to different disulfide bond configurations and long-range protein–protein interactions) must be small enough to be overcome by mechanical agitation and internal protein dynamics. The fully closed conformation (state VII) appears to be a kinetically stable conformation that accumulates over time in unagitated solutions at 4 °C. This kinetic stability may be ascribed to the dense protein packing interactions in state VII and the resulting restriction of the dynamics of both the lattice and the individual building blocks.

Geometric considerations based on the retention of $p42_12$ symmetry (Fig. 3b, d) as well as a strain analysis of the seven conformational states by digital image correlation (Fig. 3e, Extended Data Fig. 10c) indicate that ^{C98}RhuA crystals are auxetic and have a Poisson's ratio of $\nu = -1.00 \pm 0.01$ (Fig. 3f). Poisson's ratio is a scale-independent metric that describes the response of a material to strain; it is defined as the ratio between transverse (e_x) and longitudinal (e_y) strains under uniaxial loading[25]. The strains e_x and e_y are, in turn, approximated by changes in material length in the transverse (Δx) and longitudinal (Δy) directions (Fig. 3d)[25]:

$$\nu = -\frac{e_y}{e_x} \approx -\frac{\Delta y}{\Delta x}$$

where $-1 \le \nu \le 0.5$ for an isotropic 3D material and $-1 \le \nu \le 1$ for an isotropic 2D material[26]. Most materials have positive ν; that is, they become thinner in the longitudinal direction when stretched transversely[25]. By contrast, materials with negative ν (that is, auxetic materials) display the counterintuitive behaviour of longitudinal expansion upon transverse stretching. Thus, auxetic materials can be expected to possess enhanced toughness, resistance to indentation and shear stiffness as well as favourable damping and acoustic response, and have been proposed for use in protective armours, smart textiles, actuated filtration, and piezoelectric and biomedical devices[25,27–29]. Although materials with negative ν exist, most fall in the range of $-0.4 < \nu < 0$, with the lowest reported values of −0.7

© 2016 Macmillan Publishers Limited. All rights reserved

to −0.8 observed in re-entrant foams[30,31]. It has been postulated[26] that for a 2D lattice of rotating rigid squares with flexible hinges, ν should be equal to the lowest thermodynamically permissible value of −1 at all rotation angles. ^{C98}RhuA lattices represent a true realization of this theoretical model, and the first isotropic material with $\nu = -1$ designed and constructed at the molecular scale. Assuming the model of ref. 26, we calculate that ^{C98}RhuA crystals should shrink or expand simultaneously in the x and y dimensions by at least 24% during the conversion between fully open (I) and fully closed (VII) states (Extended Data Fig. 10d).

We envision that upon hierarchical assembly via physical methods or incorporation into polymeric materials using chemical strategies, molecular architectures such as 2D ^{C98}RhuA lattices may be used as feedstocks for adaptive and auxetic macroscopic materials. In general, our study underscores the utility of dynamic covalent bonds in the construction of highly ordered, yet adaptive, protein materials. Specifically, owing to their high structural quality and chemically tunable assembly under ambient conditions, ^{C98}RhuA crystals provide a unique medium for studying molecular self-assembly and crystallization as well as for investigating the energy landscape of lattice dynamics. The resultant understanding of structural dynamics at the nanoscale should greatly aid the fabrication of functional materials.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 6 October 2015; accepted 1 March 2016.
Published online 2 May 2016.

1. Mas-Ballesté, R., Gómez-Navarro, C., Gómez-Herrero, J. & Zamora, F. 2D materials: to graphene and beyond. *Nanoscale* **3,** 20–30 (2011).
2. Butler, S. Z. *et al.* Progress, challenges, and opportunities in two-dimensional materials beyond graphene. *ACS Nano* **7,** 2898–2926 (2013).
3. Schedin, F. *et al.* Detection of individual gas molecules adsorbed on graphene. *Nature Mater.* **6,** 652–655 (2007).
4. Lu, C.-H., Yang, H.-H., Zhu, C.-L., Chen, X. & Chen, G.-N. A graphene platform for sensing biomolecules. *Angew. Chem.* **121,** 4879–4881 (2009).
5. Joshi, R. K. *et al.* Precise and ultrafast molecular sieving through graphene oxide membranes. *Science* **343,** 752–754 (2014).
6. Nicolosi, V., Chhowalla, M., Kanatzidis, M. G., Strano, M. S. & Coleman, J. N. Liquid exfoliation of layered materials. *Science* **340,** 1226419 (2013).
7. Li, D., Muller, M. B., Gilje, S., Kaner, R. B. & Wallace, G. G. Processable aqueous dispersions of graphene nanosheets. *Nature Nanotechnol.* **3,** 101–105 (2008).
8. Colson, J. W. *et al.* Oriented 2D covalent organic framework thin films on single-layer graphene. *Science* **332,** 228–231 (2011).
9. Kissel, P., Murray, D. J., Wulftange, W. J., Catalano, V. J. & King, B. T. A nanoporous two-dimensional polymer by single-crystal-to-single-crystal photopolymerization. *Nature Chem.* **6,** 774–778 (2014).
10. Kory, M. J. *et al.* Gram-scale synthesis of two-dimensional polymer crystals and their structure analysis by X-ray diffraction. *Nature Chem.* **6,** 779–784 (2014).
11. Shimomura, S. *et al.* Selective sorption of oxygen and nitric oxide by an electron-donating flexible porous coordination polymer. *Nature Chem.* **2,** 633–637 (2010).
12. Rabone, J. *et al.* An adaptable peptide-based porous material. *Science* **329,** 1053–1057 (2010).
13. Serre, C. *et al.* Role of solvent-host interactions that lead to very large swelling of hybrid frameworks. *Science* **315,** 1828–1831 (2007).
14. Sleytr, U. B., Schuster, B., Egelseer, E. M. & Pum, D. S-layers: principles and applications. *FEMS Microbiol. Rev.* **38,** 823–864 (2014).
15. Hampp, N. Bacteriorhodopsin as a photochromic retinal protein for optical memories. *Chem. Rev.* **100,** 1755–1776 (2000).
16. Engel, A. *et al.* Assembly of 2-D membrane protein crystals: dynamics, crystal order, and fidelity of structure analysis by electron microscopy. *J. Struct. Biol.* **109,** 219–234 (1992).
17. Stahlberg, H. *et al.* Two-dimensional crystals: a powerful approach to assess structure, function and dynamics of membrane proteins. *FEBS Lett.* **504,** 166–172 (2001).
18. Saboe, P. O. *et al.* Two-dimensional protein crystals for solar energy conversion. *Adv. Mater.* **26,** 7064–7069 (2014).
19. Sinclair, J. C., Davies, K. M., Venien-Bryan, C. & Noble, M. E. M. Generation of protein lattices by fusing proteins with matching rotational symmetry. *Nature Nanotechnol.* **6,** 558–562 (2011).
20. Brodin, J. D. *et al.* Metal-directed, chemically tunable assembly of one-, two- and three-dimensional crystalline protein arrays. *Nature Chem.* **4,** 375–382 (2012).
21. Gonen, S., DiMaio, F., Gonen, T. & Baker, D. Design of ordered two-dimensional arrays mediated by noncovalent protein-protein interfaces. *Science* **348,** 1365–1368 (2015).
22. Lanci, C. J. *et al.* Computational design of a protein crystal. *Proc. Natl Acad. Sci. USA* **109,** 7304–7309 (2012).
23. Ringler, P. & Schulz, G. E. Self-assembly of proteins into designed networks. *Science* **302,** 106–109 (2003).
24. Grueninger, D. *et al.* Designed protein-protein association. *Science* **319,** 206–209 (2008).
25. Greaves, G. N., Greer, A. L., Lakes, R. S. & Rouxel, T. Poisson's ratio and modern materials. *Nature Mater.* **10,** 823–837 (2011).
26. Grima, J. N., Alderson, A. & Evans, K. E. Auxetic behaviour from rotating rigid units. *Phys. Status Solidi B* **242,** 561–575 (2005).
27. Evans, K. E. & Alderson, A. Auxetic materials: functional materials and structures from lateral thinking! *Adv. Mater.* **12,** 617–628 (2000).
28. Baughman, R. H. Auxetic materials: avoiding the shrink. *Nature* **425,** 667 (2003).
29. Scarpa, F., Ciffo, L. G. & Yates, J. R. Dynamic properties of high structural integrity auxetic open cell foam. *Smart Mater. Struct.* **13,** 49–56 (2004).
30. Lakes, R. Foam structures with a negative Poisson's ratio. *Science* **235,** 1038–1040 (1987).
31. Choi, J. B. & Lakes, R. S. Non-linear properties of metallic cellular materials with a negative Poisson's ratio. *J. Mater. Sci.* **27,** 5375–5381 (1992).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank M. Sailor, H. Gray, R. Alberstein and F. Paesani for discussions, N. Olson and J. Bouwer for assistance with TEM measurements, and N. Gianneschi for the use of AFM and DLS instrumentation. This work was primarily supported by the US Department of Energy (DOE) (Division of Materials Sciences, Office of Basic Energy Sciences, Award DE-FG02-10ER46677 to F.A.T.). Y.S. and F.A.T. were partially supported (for EM measurements) by a grant from the Air Force Office of Scientific Research (AFOSR, BRI FA9550-12-1-0414). The EM facilities used in this work are supported by funding to T.S.B. from the NIH (1S10 RR-020016 and GM-033050), the Agouron Foundation and UCSD.

Author Contributions Y.S. and F.A.T. conceived the work and designed the experiments. Y.S. carried out all experimental work and EM measurements. G.C. and Y.S. performed, and T.S.B. oversaw, EM data analysis. D.R. and P.D.Z. performed digital image-correlation analysis. F.A.T. and Y.S. wrote the manuscript with contributions from G.C., T.S.B., D.R. and P.D.Z.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to F.A.T. (tezcan@ucsd.edu).

© 2016 Macmillan Publishers Limited. All rights reserved

METHODS

Design of RhuA variants and site-directed mutagenesis. RhuA variants were designed as described in the main text on the basis of previously reported crystal structures (Protein Data Bank (PDB) 1GT7[23] for ^{C98}RhuA, $^{H63/H98}$RhuA, ^{D98}RhuA, ^{C133}RhuA and ^{C266}RhuA, and PDB 2UYU[24] for $^{F88/C98}$RhuA. All RhuA variants also contain the mutations E192A and C126S as previously reported[23,24]. The gene for ^{C98}RhuA, pre-inserted into the pJ414 expression vector optimized for expression in *E. coli*, was purchased from DNA2.0. $^{F88/C98}$RhuA, $^{H63/H98}$RhuA, ^{D98}RhuA, ^{C133}RhuA and ^{C266}RhuA were prepared using QuikChange mutagenesis (Stratagene) with primers obtained from Integrated DNA Technologies (Supplementary Table 1). The mutant plasmids were transformed into XL-1 Blue *E. coli* cells followed by purification using the QIAprep Spin Miniprep kit (Qiagen). The presence of the mutations was verified by sequencing (Retrogen). Amino acid sequences of RhuA variants are shown in Supplementary Table 2.

Protein expression and purification. Bacterial expression of RhuA variants was performed according to previously published procedures with slight modifications[32]. Plasmids bearing the variants were transformed into BL21(DE3) *E. coli* cells, and the colonies were grown overnight at 37 °C on lysogeny broth (LB) agar plates (pH 7.4) containing 100 mg l^{-1} ampicillin. Starter cultures (5 ml with 100 mg l^{-1} ampicillin) from single colonies were grown for about 4–6 h at 37 °C (with shaking at 250 r.p.m.) before inoculation into 1-l LB cultures containing 100 mg l^{-1} ampicillin. After the cells were grown to an optical density of about 0.8 at 600 nm, protein expression was induced with 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG, Gold Biotechnology) for 12–13 h (at 37 °C with shaking at 250 r.p.m.). Cells were pelleted by centrifugation (5,000 r.p.m. at 4 °C for 10 min), and resuspended in a buffer solution containing 10 mM tris(hydroxymethyl)aminomethane hydrochloride (Tris; pH 7.5), 1 mM $ZnCl_2$ and 10 mM β-mercaptoethanol (βME). For $^{H63/H98}$RhuA preparations, $ZnCl_2$ was excluded from buffer solutions to avoid possible protein precipitation during purification. Cell lysis was performed by sonication for 15 min on ice. The lysis solution was centrifuged for 30 min at 12,000 r.p.m. at 4 °C. Polymin-P (Acros) was added to the supernatant at a final concentration of 0.15% (w/v) for nucleic acid precipitation and the resulting mixture was stirred for 30 min before centrifugation for 30 min at 12,000 r.p.m. at 4 °C. The supernatant was loaded onto a DEAE-Sepharose CL-6B (GE Healthcare) column and eluted using a gradient of 0–500 mM NaCl in Tris buffer at 4 °C. Fractions containing RhuA, which eluted at approximately 200 mM NaCl, were added to 1.7 M ammonium sulfate, and were centrifuged for 45 min at 12,000 r.p.m. at 4 °C. The precipitate was dissolved in a buffer solution containing 5 mM sodium phosphate (NaPi; pH 7.2), 1 mM $ZnCl_2$ and 10 mM βME, and then dialysed three times against 5 l of the same buffer solution. Further purification was done using either a High Q cartridge column (BioRad; at pH 8.0) using a 0–500 mM NaCl gradient or a Mini CHT Type I hydroxyapatite column (BioRad; at pH 7.2) using a 5–500 mM NaPi gradient on a DuoFlow fast protein chromatography workstation (BioRad). Pure protein fractions eluted at 350 mM NaCl and 200 mM NaPi, respectively. These fractions were combined, dialysed three times against a buffer solution of 10 mM Tris (pH 7.5), 1 mM $ZnCl_2$ and 10 mM βME, and concentrated in an Amicon stirred cell at 4 °C. The protein was then flash frozen and kept at −80 °C until use. Protein purity was confirmed by SDS PAGE (Supplementary Fig. 1) and ESI-MS (Supplementary Fig. 2 and Supplementary Table 3), and stock concentrations were determined as previously reported[33].

Preparation of 2D RhuA crystals. All 2D RhuA crystals reported here were obtained in an unsupported fashion in solutions. As described in the main text, several solution conditions were screened for optimizing the formation of 2D protein crystals of ^{C98}RhuA, $^{F88/C98}$RhuA and $^{H63/H98}$RhuA variants. For all variants, these screening conditions included various pH values (6–8), starting protein concentrations (25–175 μM), buffer salt concentrations and identities (5–20 mM bis-Tris, CHES, MES, MOPS, NaPi and Tris) and temperatures (4–37 °C). For the disulfide-mediated self-assembly of ^{C98}RhuA and $^{F88/C98}$RhuA, we additionally screened various oxidizing conditions (1–2 mM total GSH/GSSG at ratios varying from 1:19 to 19:1, or 5–10 mM βME, which slowly decomposes in aqueous solution). For the metal-mediated assembly of $^{H63/H98}$RhuA, we screened 4–20 molar equivalents (over tetrameric protein concentration) of $ZnCl_2$ and $CuCl_2$. As stated in the main text, the best conditions were: ^{C98}RhuA and $^{F88/C98}$RhuA (≥125 μM protein, 10 mM βME, pH 7.5, 10 mM Tris); $^{H63/H98}$RhuA (25 μM protein, 200 μM $ZnCl_2$, pH 7, 20 mM MOPS); all at 4 °C. In general, the formation of crystals was either immediate (upon metal addition for $^{H63/H98}$RhuA) or lasted overnight to several days (for ^{C98}RhuA and $^{F88/C98}$RhuA). Crystal formation resulted in increasing cloudiness of the self-assembly solutions. The growth of ^{C98}RhuA crystals could be accelerated by gentle shaking (Extended Data Fig. 2a, b). In additional experiments, we tested the thermo- and chemo-stability of ^{C98}RhuA crystals, as shown in Supplementary Fig. 3.

Electron microscopy. For negative-stain TEM sample preparation, 3–3.5-μl aliquots of crystal suspensions were applied onto negatively glow-discharged carbon-coated Cu grids (Ted Pella, Inc.), washed with Milli-Q water, and stained with 1% uranyl acetate at 4 °C. For cryo-EM sample preparation, 3–3.5-μl aliquots of 25 μM ^{C98}RhuA sample (diluted sample) were deposited onto negatively glow-discharged Quantifoil grids, and then plunged into liquid ethane after blotting. The sample was then stored under liquid nitrogen until analysis. Sample screening was performed in an FEI Sphera transmission electron microscope equipped with a LaB_6 electron gun at 200 keV and imaged on Gatan 2K^2 CCD. A complete electron crystallographic analysis was done on 43, 50 and 33 micrographs for ^{C98}RhuA, $^{H63/H98}$RhuA and $^{F88/C98}$RhuA, respectively, which were selected by visual assessment. Micrographs were each processed separately using the 2dx software[33], which implements a semi-automatic processing pipeline that is mostly based on programs from the MRC suite[34]. The program allows one to determine the 2D plane group lattice symmetry of each crystal, to calculate its unit-cell dimensions and to generate a projection density map. The processing involves estimation of the contrast transfer function (CTF) using CTFFIND3[35], spot-list determination, automatic lattice determination, crystal masking, unbending and generation of a projection map after CTF correction. Within this program, the plane group symmetry is determined using ALLSPACE[36] by comparing the internal phase residual for each symmetry to its expected value. Lattice unit-cell dimensions were determined on a subset of images, screened by the quality of the reflections—12, 44 and 13 for ^{C98}RhuA, $^{H63/H98}$RhuA and $^{F88/C98}$RhuA, respectively. The estimated resolution limit of each image (15–30 Å) was assessed by both visual inspection of its computed Fourier transform and comparison with simulated images filtered at different levels of resolution.

AFM measurements. For AFM sample preparation, 2 μl of crystal suspensions were applied on freshly cleaved mica, washed with deionized water and dried. AFM images were collected using a Dimension Icon microscope (Bruker) using Scan Asyst peak force tapping (in air) mode at a resolution of 512 lines per image and a scan rate of 0.3–0.5 Hz. The AFM images obtained were processed using Nanoscope Analysis (Bruker).

SEM measurements. For SEM sample preparation, 2 μl of crystal suspensions were applied onto SiO_2 substrates. The sample was sputter-coated with iridium oxide before analysis to reduce charging effects. SEM micrographs were collected on an FEIXL 30UHR scanning electron microscope operated at 5 keV.

DLS measurements. DLS experiments were performed using a Wyatt DynaPro NanoStar instrument. The experiment runs were performed to collect 10 acquisitions with 657-nm excitation at a power setting of 100%. Measurements were plotted using a Rayleigh sphere model. Peak radius cut-offs were fixed according to default settings (0.5–10,000 nm).

Structural modelling and simulations. Projected electron-density maps were used as a reference in the visualization software UCSF Chimera[37] to determine the orientation of the RhuA molecules in each crystal. The atomic coordinates were extracted from PDB entries 1GT7 (C_4-symmetric tetramer for ^{C98}RhuA and $^{H63/H98}$RhuA) and 2UYU (D_4-symmetric octamer for $^{F88/C98}$RhuA) and placed manually to fit the position of one subunit in the observed projected density map. The orientation of the molecule was refined in an iterative manner by comparing the experimental density with the density simulated from a model of the unit cell, obtained as explained below. Simulated projected electron-density maps were computed using Bsoft (http://lsbr.niams.nih.gov/bsoft)[38] and EMAN2 (http://blake.bcm.edu/emanwiki/EMAN2)[39]. Starting from the estimated position and orientation of a single subunit, multiple copies of the model were generated to cover one unit cell (*bmoledit* in Bsoft) and converted to electron density (*e2pdb2mrc.py* in EMAN2) with a resolution limit of 30 Å. Finally, the density volume was projected along the vertical direction (*bproject* in Bsoft) to generate a 2D density map comparable to the experimental map.

Classification of the conformational states of 2D ^{C98}RhuA crystals. TEM micrographs of ^{C98}RhuA crystals were analysed computationally to categorize different conformational states of the 2D lattice. Essentially, for each raw image we determined a 'roundness index' number derived from the shape of the pores in the lattice, and used this value to classify the images. The analysis was performed on 982 images using an ad hoc algorithm coded as a macro in Fiji (http://fiji.sc/Fiji)[40] and ImageJ (http://imagej.nih.gov/ij)[41]. Each image was initially processed by applying a 3 × 3 mean filter three times, and then converted to a binary image after automatic thresholding to segment out the pores. The mask was then filtered using the opening morphological operator to smooth the shape of the pores and to remove small outliers. Pores that were distorted from potential bending of the crystal in some regions of the field of view were discarded using their size for screening. For images acquired at nominal magnifications of ×50,000 and ×68,000 (calibrated pixel size of 2.033 Å and 1.646 Å, respectively), only pores with areas that covered 500–1,200 pixels were kept. The remaining pores were modelled as ellipses and their roundness index was determined as the ratio between the lengths of the major and the minor elliptical axes. A single index value was assigned to each micrograph as the average of all the segmented pores. The images were assigned

© 2016 Macmillan Publishers Limited. All rights reserved

to one of the seven identified states using the following criterion: state I, >0.85; state II, 0.75–0.85; state III, 0.65–0.75; state IV, 0.55–0.65; state V, 0.45–0.55; state VI, 0.35–0.45; state VII, <0.35.

Generation of video simulating the lattice motions. Supplementary Video 1 was generated to illustrate the dynamics of 2D ^{C98}RhuA crystals starting from the TEM snapshots of the seven conformational states (I–VII; Fig. 3a). All projected density images were rescaled to include the same field of view (*bint* and *bimg* from Bsoft). Twenty-four intermediate image frames were created between each pair in the sequence by pseudo-morphing, using the program *convert* from the software suite ImageMagick. Finally, the video was generated with *ffmpeg* (http://www.ffmpeg.org) by combining the images as frames.

Digital image correlation for determining Poisson's value. Using Matlab, we performed digital image processing of the reconstructed 2D TEM images of dynamic ^{C98}RhuA crystals for evaluating Poisson's ratios. As a first step, histogram normalization was performed, taking the image of conformational state I as a reference to normalize the intensity of the images (Extended Data Fig. 10c). This normalization guarantees that the thresholds used for pixel selection have the same meaning for all the images. After normalizing, we proceeded to select the representative volume element (RVE) in each of the images. The RVE considered has a rectangular section whose vertices correspond to the centroid of the lattice pores (shown as green rectangles in Extended Data Fig. 10c and red rectangles in Fig. 3e). To find the position of the centroids, we determined the edges of the pores using the Sobel edge-detection method. The edges of the pores are shown as purple lines in Extended Data Fig. 10c and Fig. 3e. Then, centroid calculation was performed by taking the mean of the coordinates in x and y for each pixel in the border. After the selection of the RVE, the size of the RhuA building blocks was measured on each image by defining the circle that is circumscribed to each square (shown as blue circles in Extended Data Fig. 10c). Because the square shape of the RhuA building blocks can be assumed to remain rigid in each conformational state, we determined that the images were at slightly different scales; thus, appropriate magnification factors were calculated taking the image corresponding to conformational state I as a reference. The applied scaling factors are shown in Extended Data Fig. 10c. We then calculated the position of the vertices of the RVE on each conformational state and proceeded with the calculation of Poisson's ratio. For each RVE, local values of engineering strains e_x and e_y are calculated from the vectors $\boldsymbol{M}$, $\boldsymbol{N}$ and $\boldsymbol{m}$, $\boldsymbol{n}$ that span the RVE on its reference (state I) and deformed configurations (states II–VII), respectively (see Fig. 3e):

$$\boldsymbol{M}=\begin{bmatrix}M_x\\M_y\end{bmatrix},\quad \boldsymbol{N}=\begin{bmatrix}N_x\\N_y\end{bmatrix};\qquad \boldsymbol{m}=\begin{bmatrix}m_x\\m_y\end{bmatrix},\quad \boldsymbol{n}=\begin{bmatrix}n_x\\n_y\end{bmatrix}$$

The deformed and undeformed configurations are related through the deformation gradient F:

$$\begin{bmatrix}m_x\\m_y\end{bmatrix}=F\begin{bmatrix}M_x\\M_y\end{bmatrix};\quad \begin{bmatrix}n_x\\n_y\end{bmatrix}=F\begin{bmatrix}N_x\\N_y\end{bmatrix}$$

where:

$$F=\begin{bmatrix}1+e_x & e_{xy}\\ e_{yx} & 1+e_y\end{bmatrix}$$

Homogenized values of the engineering strains for the RVE are calculated as:

$$\begin{bmatrix}1+e_x & e_{xy}\\ e_{yx} & 1+e_y\end{bmatrix}=\begin{bmatrix}n_x & m_x\\ n_y & m_y\end{bmatrix}\begin{bmatrix}N_x & M_x\\ N_y & M_y\end{bmatrix}^{-1}$$

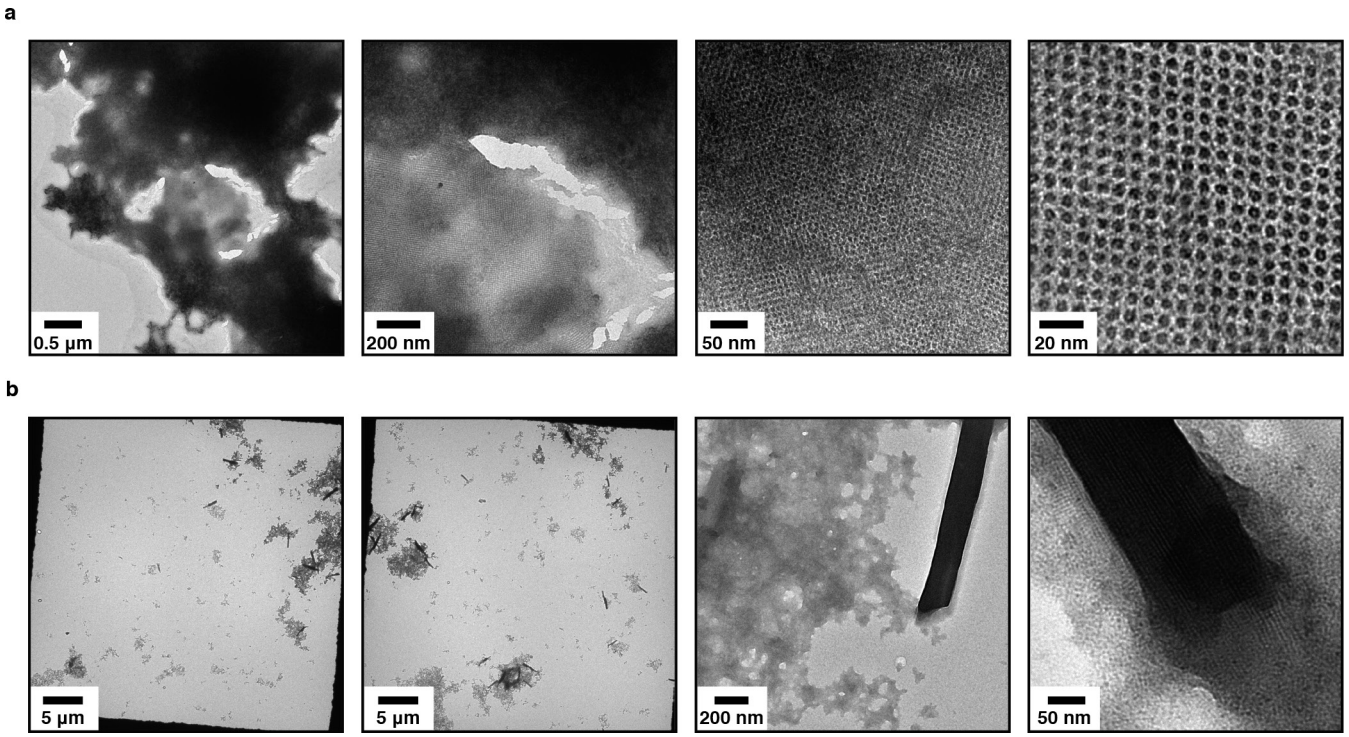
Finally, Poisson's ratio is calculated as:

$$\nu = -\frac{e_y}{e_x}$$

To determine the uncertainties in ν (error bars in Fig. 3f), we calculated the local standard deviation given the distribution of pixel intensity of each Fourier transform image (using the command *stdfilt* in Matlab). From this local standard deviation, we obtained the uncertainty in the intensity of each pixel that forms the edges of the pores. These uncertainties were propagated in calculating the positions of RVE vertices and finally in calculating the Poisson's ratios.

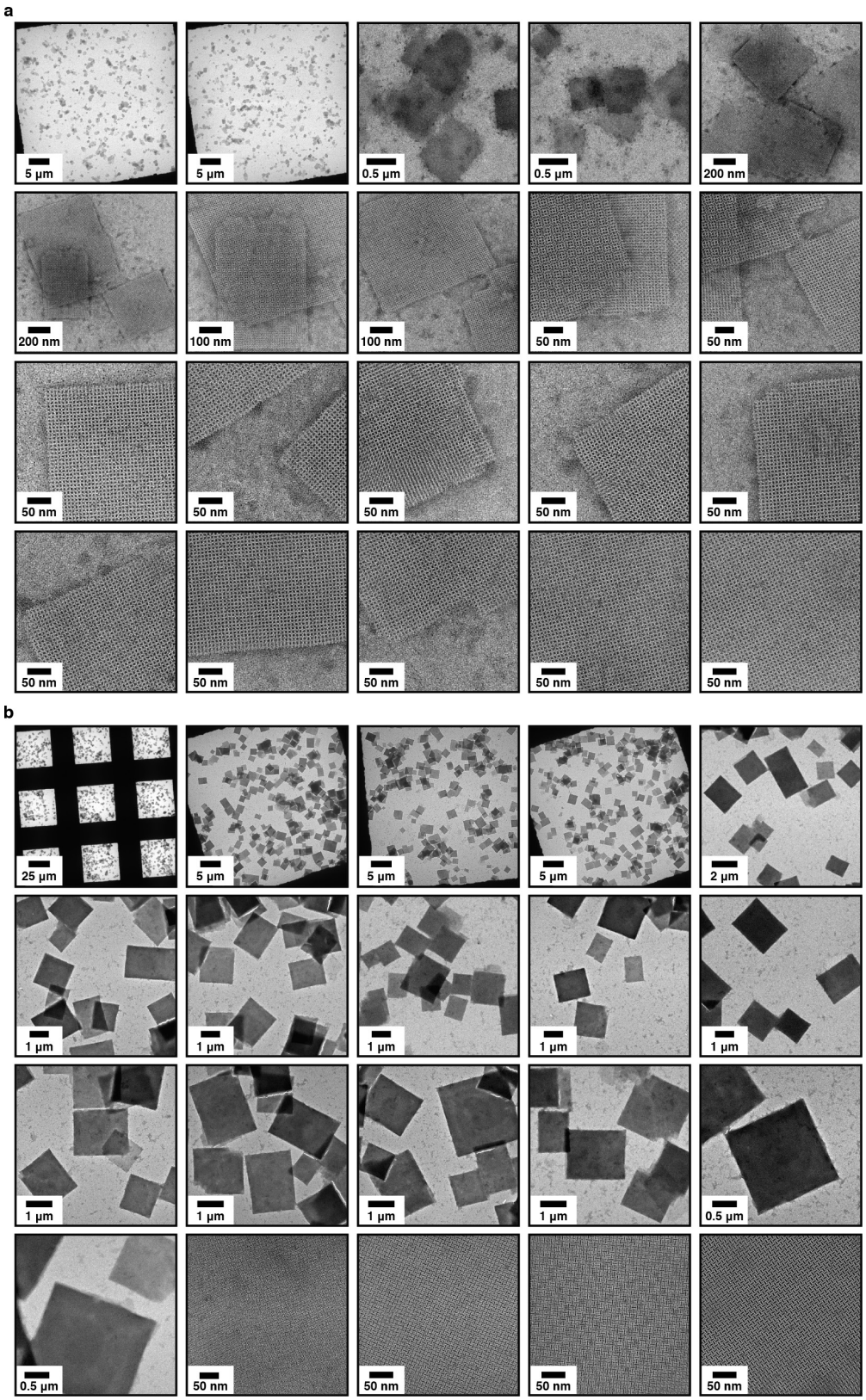
32. Kroemer, M. & Schulz, G. E. The structure of L-rhamnulose-1-phosphate aldolase (class II) solved by low-resolution SIR phasing and 20-fold NCS averaging. *Acta Crystallogr.* **D58,** 824–832 (2002).
33. Edelhoch, H. Spectroscopic determination of tryptophan and tyrosine in proteins. *Biochemistry* **6,** 1948–1954 (1967).
34. Crowther, R., Henderson, R. & Smith, J. MRC image processing programs. *J. Struct. Biol.* **116,** 9–16 (1996).
35. Mindell, J. A. & Grigorieff, N. Accurate determination of local defocus and specimen tilt in electron microscopy. *J. Struct. Biol.* **142,** 334–347 (2003).
36. Valpuesta, J. M., Carrascosa, J. L. & Henderson, R. Analysis of electron microscope images and electron diffraction patterns of thin crystals of Ø29 connectors in ice. *J. Mol. Biol.* **240,** 281–287 (1994).
37. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).
38. Heymann, J. B. Bsoft: image and molecular processing in electron microscopy. *J. Struct. Biol.* **133,** 156–169 (2001).
39. Tang, G. *et al.* EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157,** 38–46 (2007).
40. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nature Methods* **9,** 676–682 (2012).
41. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nature Methods* **9,** 671–675 (2012).
42. Henderson, R., Baldwin, J. M., Downing, K. H., Lepault, J. & Zemlin, F. Structure of purple membrane from halobacterium halobium: recording, measurement and evaluation of electron micrographs at 3.5 Å resolution. *Ultramicroscopy* **19,** 147–178 (1986).

© 2016 Macmillan Publishers Limited. All rights reserved



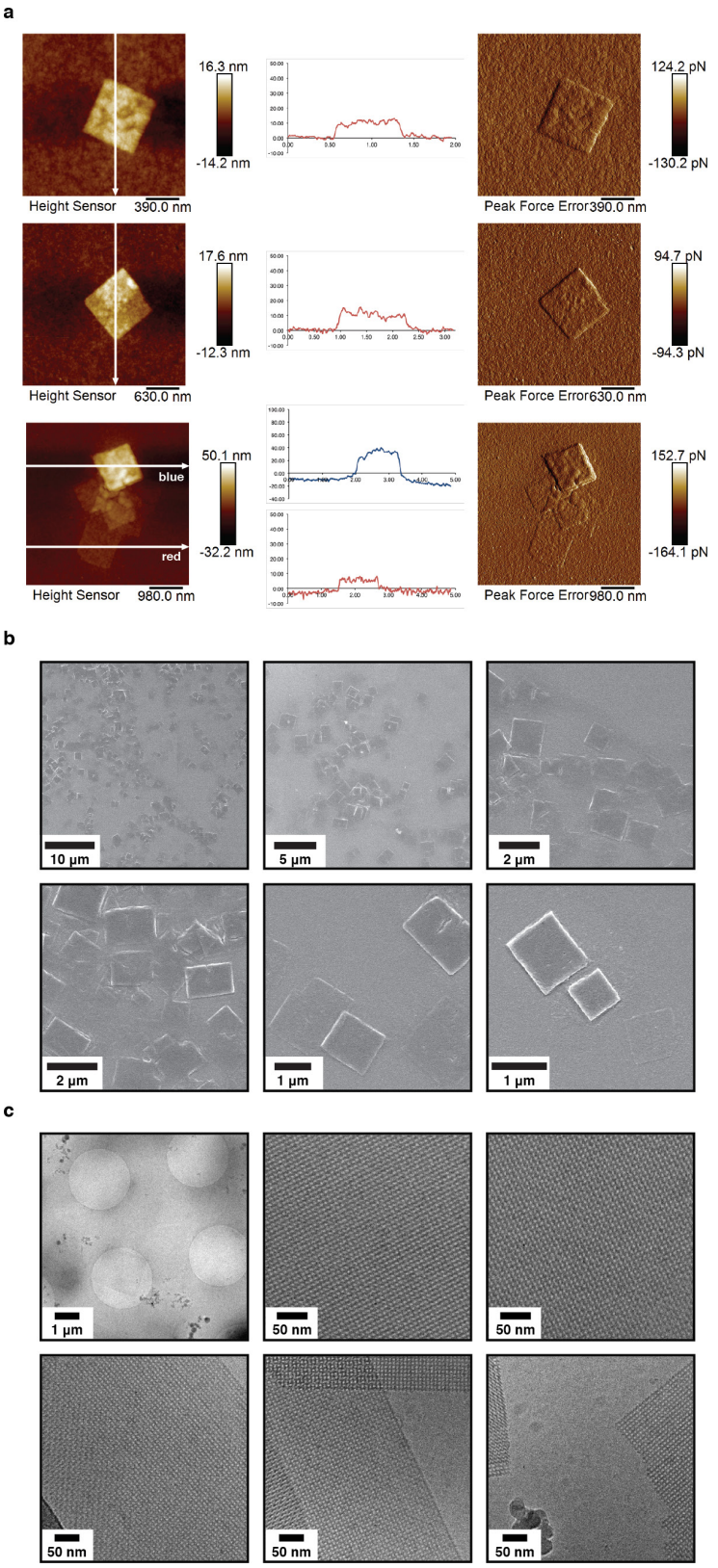
Extended Data Figure 1 | TEM characterization of ^{C98}RhuA and $^{H63/H98}$RhuA self-assembly under non-optimal conditions. a, 25 µM ^{C98}RhuA was incubated at 4 °C under air exposure for 1 day; the solution contained 20 mM Tris (pH 8) and no reductants or oxidants. **b**, 25 µM $^{H63/H98}$RhuA was incubated in the presence of 1 mM $ZnCl_2$ at 4 °C for 1 day; the solution contained 20 mM MOPS (pH 7).

© 2016 Macmillan Publishers Limited. All rights reserved



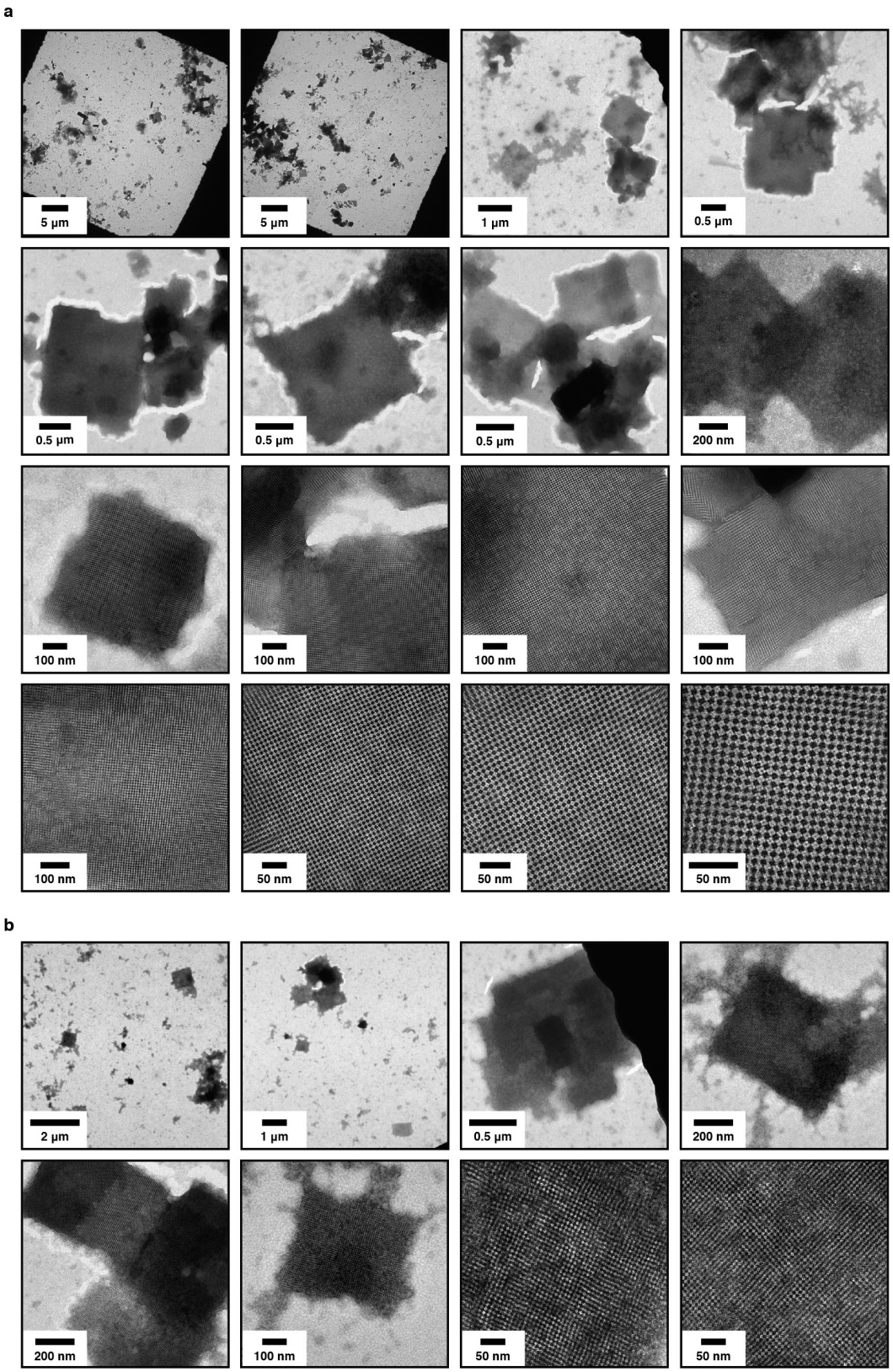
Extended Data Figure 2 | TEM characterization of optimized 2D ^{C98}RhuA crystals. a, 125 µM ^{C98}RhuA was incubated in the presence of 10 mM βME at 4 °C in a standing solution for 3 days. **b**, 125 µM ^{C98}RhuA was incubated in the presence of 10 mM βME at 4 °C with gentle shaking for 3 days, followed by 2 days at rest, during which a dense precipitate of crystals formed.

© 2016 Macmillan Publishers Limited. All rights reserved



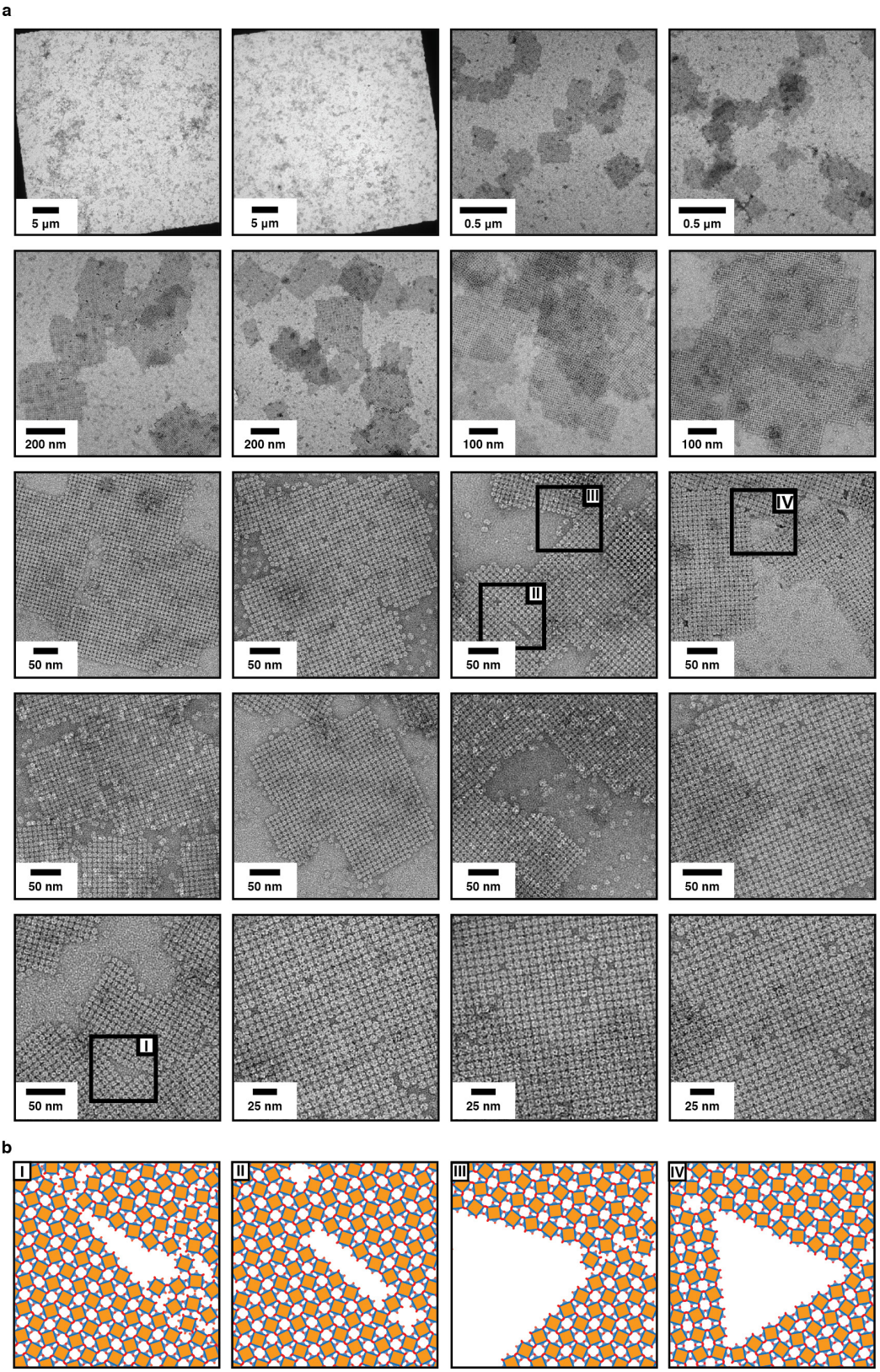
Extended Data Figure 3 | Additional structural characterization of 2D ^{C98}RhuA crystals. a, AFM; **b**, SEM; **c**, Cryo-TEM. The middle column in **a** shows the profiles along the arrows in the left column.

© 2016 Macmillan Publishers Limited. All rights reserved



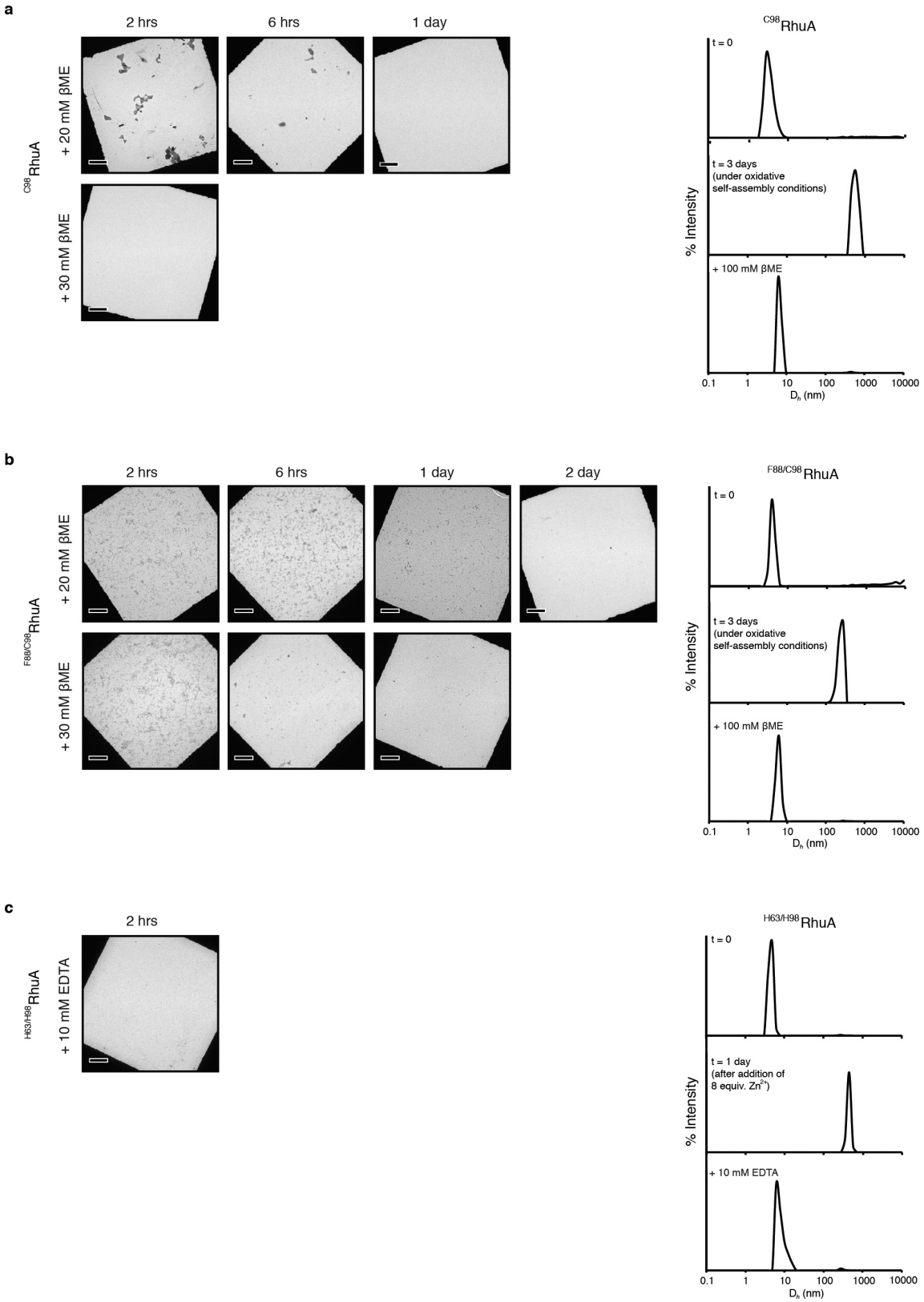
Extended Data Figure 4 | TEM characterization of 2D $^{H63/H98}$RhuA crystals. a, **b**, 25 μM $^{H63/H98}$RhuA was incubated with 200 μM $ZnCl_2$ (**a**) or $CuCl_2$ (**b**) in a 20 mM MOPS buffer solution (pH 7.0) at 4 °C for 1 day.

© 2016 Macmillan Publishers Limited. All rights reserved



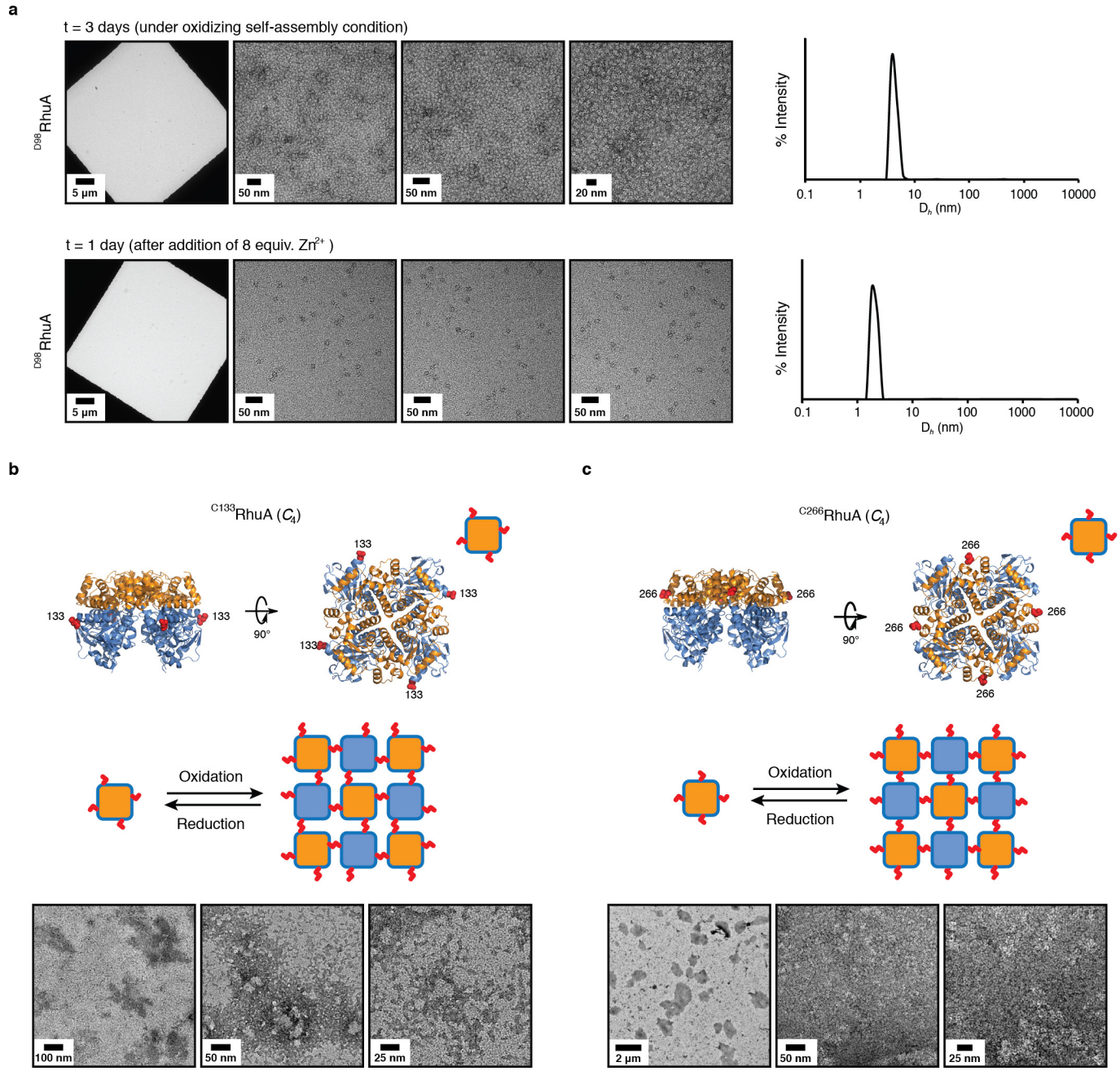
Extended Data Figure 5 | TEM characterization of 2D $^{F88/C98}$RhuA crystals. a, 125 µM $^{F88/C98}$RhuA was incubated in the presence of 10 mM βME at 4 °C with gentle shaking for 2 days. **b**, Schematic representations of various types of defects observed in $^{F88/C98}$RhuA crystals, corresponding to boxed areas (I–IV) in **a** (colouring as in Fig. 1).

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 6 | Reversibility of oxidative or metal-mediated self-assembly of RhuA variants. a, **b**, ^{C98}RhuA (**a**) and $^{F88/C98}$RhuA (**b**) crystals. Left panels, the crystals were incubated in the presence of 20 mM or 30 mM βME (as indicated) and imaged by TEM at the indicated times after addition of βME. Right panels, 125 μM ^{C98}RhuA (**a**) or $^{F88/C98}$RhuA (**b**) were incubated under oxidative self-assembly conditions (in the presence of 10 mM βME at 4 °C with gentle shaking) and self-assembly was monitored by DLS at $t = 0$ (top), $t = 3$ days (centre) and upon addition of 100 mM βME after self-assembly (bottom). D_h refers to the hydrodynamic diameter. **c**, Left panel, $^{H63/H98}$RhuA crystals were incubated in the presence of 10 mM EDTA and imaged by TEM at 2 h after addition of EDTA. Right panels, 25 μM $^{H63/H98}$RhuA was incubated under metal-mediated self-assembly conditions (in the presence of 200 μM $ZnCl_2$ at 4 °C) and self-assembly was monitored by DLS at $t = 0$ (top), $t = 3$ days (centre) and upon addition of 10 mM EDTA (bottom). Scale bars are 5 μm in all panels.

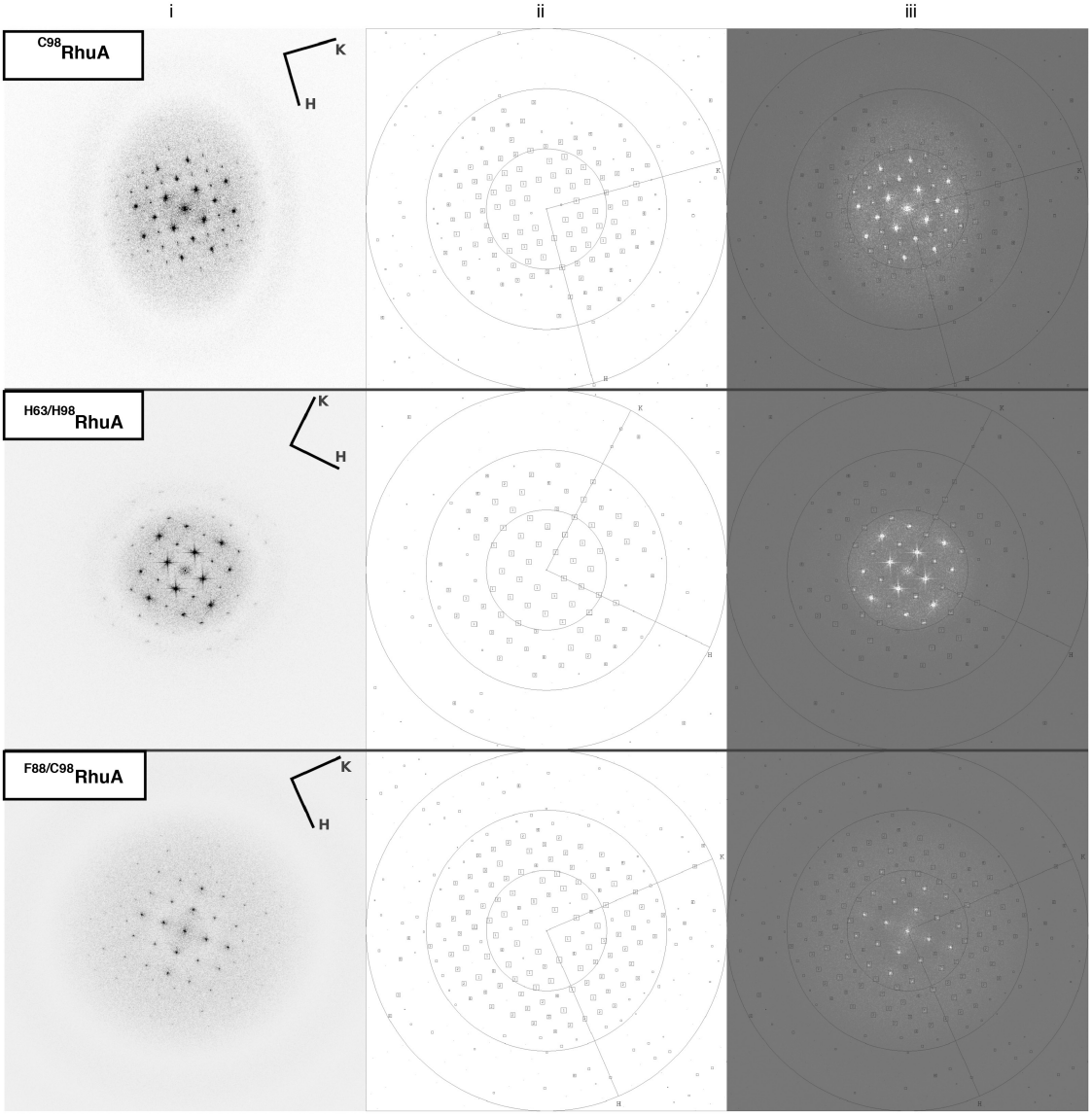
© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 7 | Characterization of the self-assembly of ^{D98}RhuA, ^{C133}RhuA and ^{C266}RhuA variants by TEM and DLS. a, TEM (right panels) and DLS (left panels) characterization of ^{D98}RhuA self-assembly under oxidative or metal-mediated self-assembly conditions that were optimized for ^{C98}RhuA; see main text or Extended Data Figs 2 and 4 for details. **b**, Possible mode of disulfide-mediated self-assembly of ^{C133}RhuA (top panels) and TEM characterization of the self-assembly products obtained under conditions that were optimized for ^{C98}RhuA (bottom panels). **c**, Possible mode of disulfide-mediated self-assembly of ^{C266}RhuA (top panels) and TEM characterization of the self-assembly products obtained under conditions that were optimized for ^{C98}RhuA (bottom panels). No crystalline assemblies were detected under these conditions for any of these three variants (or other conditions that were used for screening ^{C98}RhuA self-assembly). Colouring in the top panels of **b** and **c** as in Fig. 1.

© 2016 Macmillan Publishers Limited. All rights reserved

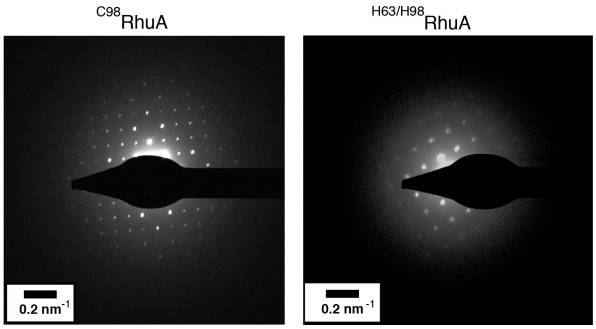
a



b

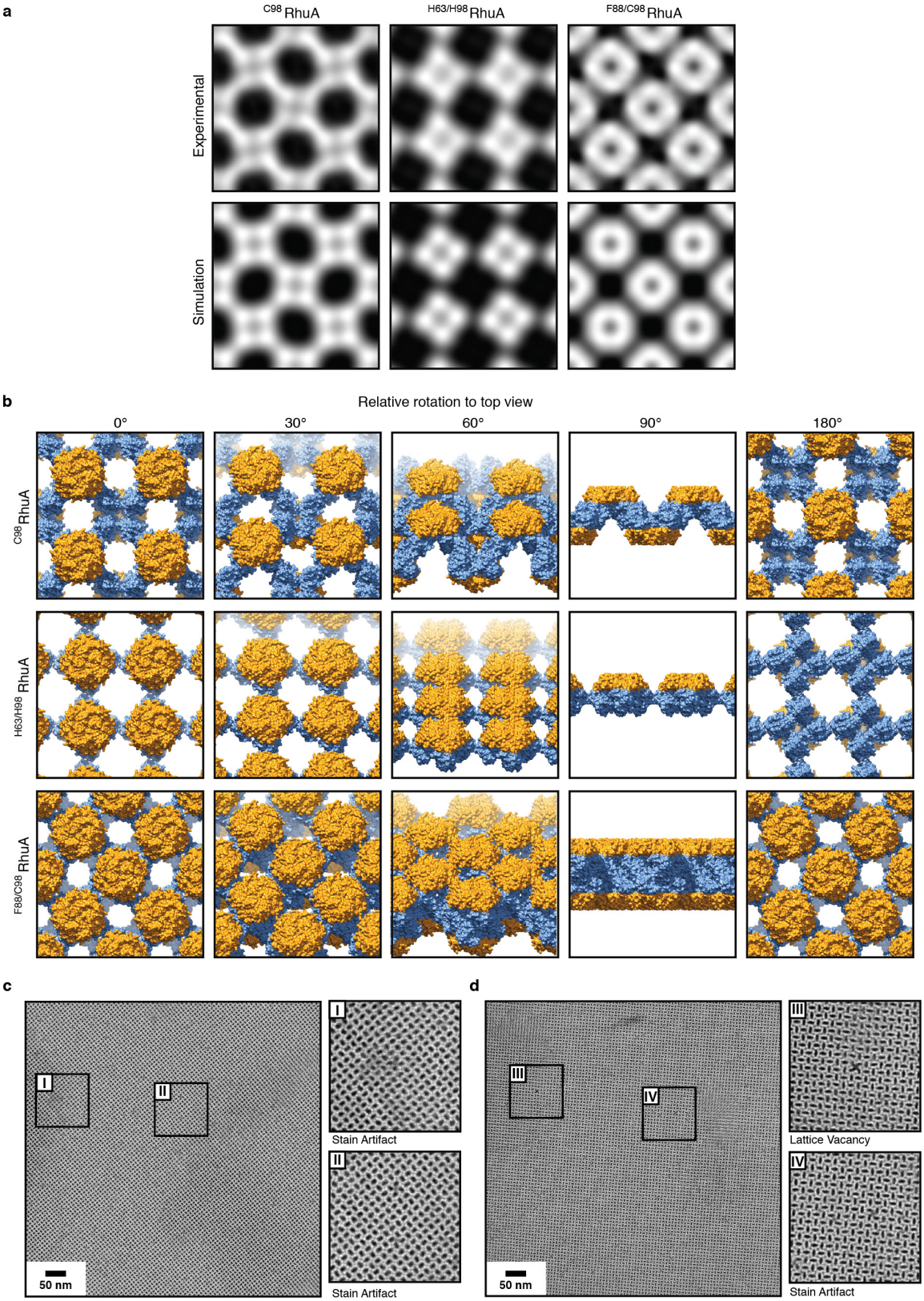
	C98 RhuA	H63/H98 RhuA	F88/C98 RhuA
No. of images used for analysis	43 (12)	50 (44)	33 (13)
Unit cell dimensions			
a = b (Å)	114 ± 3	91 ± 3	115 ± 1
γ (°)	90	90	90
Plane-group symmetry	$p42_12$	p4	$p42_12$

c



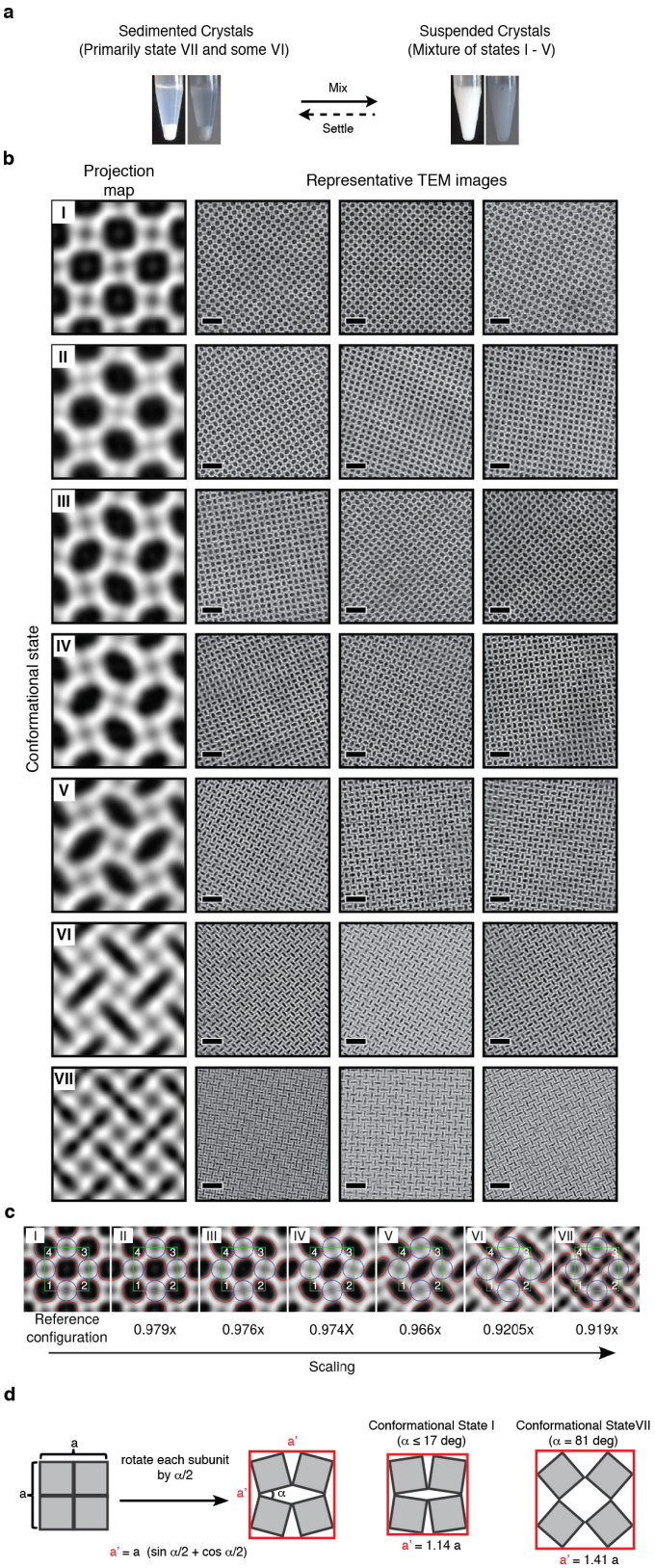
Extended Data Figure 8 | Crystallographic analysis of the 2D lattices of RhuA variants. a, (i) Representative Fourier transforms calculated from the full field of view of crystal images, displayed up to 10 Å. Reciprocal lattice axes are indicated with H and K. The reflections are consistent with the plane group symmetries estimated from the analysis of the phase residuals (see Supplementary Table 4). (ii) Integer quality (IQ) plots calculated from the spectra in column (i). The size of the boxes around the reflections reflects their IQ value, defined as the ratio of the reflection-peak amplitude and the amplitude of the background signal around each peak[42]. The most significant reflections are labelled with their IQ values, 1–4. Rings are displayed at resolutions of 30 Å, 15 Å and 10 Å. (iii) Overlap of fast Fourier transform and IQ plots. **b**, Crystallographic data for the 2D lattices of RhuA variants. Reported are the numbers of TEM images used for the analysis of plane group symmetry. The numbers in parentheses are the number of images used for determining the statistics for the unit-cell dimensions of each lattice. The plane group symmetry was determined by consensus from all the available images. **c**, Electron diffraction patterns of C98RhuA and H63/H98RhuA lattices.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 9 | 2D TEM reconstruction and structural modelling of the 2D lattices of RhuA variants. a, Comparison of the observed (experimental) and simulated projection maps of the 2D lattices of RhuA variants. **b**, Molecular arrangements in ^{C98}RhuA, $^{H63/H98}$RhuA and $^{F88/C98}$RhuA lattices viewed from different angles (colouring as in Fig. 1). **c**, Representative images of 2D ^{C98}RhuA crystals containing no defects. **d**, Representative images of 2D ^{C98}RhuA crystals containing a single lattice vacancy. The stain artefact images are shown to highlight areas that may appear to contain lattice vacancies in low-magnification views, but actually do not.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 10 | Conformational analysis of ^{C98}RhuA lattices. a, Photographs of sedimented and resuspended ^{C98}RhuA crystals (alternative views with different backgrounds are shown for clarity). **b**, Representative TEM images for different conformational states (I–VII) of the ^{C98}RhuA crystals shown in Fig. 3. Scale bars are 25 nm in all panels. **c**, Digital image processing of reconstructed images of each conformational state. Red lines represent the border of the pores in the lattice structure. Selected RVEs are shown with green lines, with the vertices of the RVEs numbered 1–4. Blue lines show the circles circumscribed to the square RhuA building block. Values of the scaling between these and the TEM images (with conformational state I as a reference) are shown below each image. **d**, Changes in the dimensions of a ^{C98}RhuA lattice of arbitrary size assuming a rotating-rigid-squares model.

© 2016 Macmillan Publishers Limited. All rights reserved

LETTER

doi:10.1038/nature17432

Iron(III)-catalysed carbonyl–olefin metathesis

Jacob R. Ludwig[1], Paul M. Zimmerman[1], Joseph B. Gianino[1] & Corinna S. Schindler[1]

The olefin metathesis reaction of two unsaturated substrates is one of the most powerful carbon–carbon-bond-forming reactions in organic chemistry. Specifically, the catalytic olefin metathesis reaction has led to profound developments in the synthesis of molecules relevant to the petroleum, materials, agricultural and pharmaceutical industries[1]. These reactions are characterized by their use of discrete metal alkylidene catalysts that operate via a well-established mechanism[2]. While the corresponding carbonyl–olefin metathesis reaction can also be used to construct carbon–carbon bonds, currently available methods are scarce and severely hampered by either harsh reaction conditions or the required use of stoichiometric transition metals as reagents. To date, no general protocol for catalytic carbonyl–olefin metathesis has been reported. Here we demonstrate a catalytic carbonyl–olefin ring-closing metathesis reaction that uses iron, an Earth-abundant and environmentally benign transition metal, as a catalyst. This transformation accommodates a variety of substrates and is distinguished by its operational simplicity, mild reaction conditions, high functional-group tolerance, and amenability to gram-scale synthesis. We anticipate that these characteristics, coupled with the efficiency of this reaction, will allow for further advances in areas that have historically been enhanced by olefin metathesis.

Classified by their mechanistic profile, metathesis reactions most often proceed through a [2+2] cycloaddition to form a 4-atom cyclic intermediate, followed by [2+2] cycloreversion, resulting in an exchange of reactive partners[3–6]. Included in this mode of reactivity are classical olefination reactions (Fig. 1a), in which highly polarized nucleophilic components add to carbonyl derivatives to give cyclic intermediates[7–9]. These compounds then fragment to produce olefins and a metal or main-group oxo species, the latter being the thermodynamic driving force for the reaction. The utility of this chemistry is often hampered by harsh conditions, the need for pre-functionalized intermediates and the production of excess quantities of waste products. A revolutionary development in this regard was the establishment of the catalytic olefin–olefin metathesis reaction, in which transition-metal alkylidenes facilitate [2+2] cycloadditions and subsequent [2+2] cycloreversions from simple olefin starting materials (Fig. 1b). Since its discovery, olefin metathesis has led to marked advances in the fields of organic chemistry as well as chemical biology, and allowed entry into previously inaccessible complex molecules. Although it was discovered at around the same time, the corresponding carbonyl–olefin metathesis (Fig. 1c) is much less developed[10]. To date, no catalytic protocol of general synthetic utility exists for carbonyl–olefin metathesis reactions, and, furthermore, available procedures

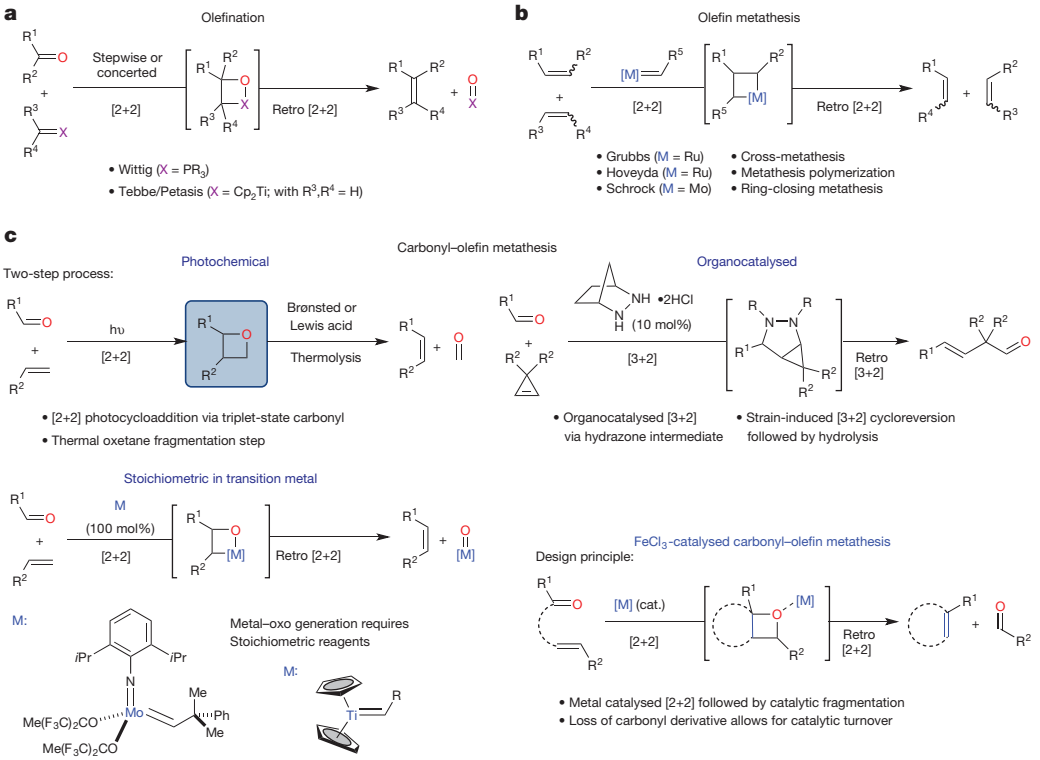


Figure 1 | Olefination and metathesis reactions for the formation of alkenes. a, Summary of traditional olefination reactions in organic synthesis. **b**, Olefin–olefin metathesis catalysed by ruthenium, molybdenum and tungsten complexes. **c**, Modes of carbonyl–olefin metathesis, including two-step photochemical cycloaddition/thermolysis cycloreversion, stoichiometric transition-metal-mediated, organocatalytic, and the new iron(III)-catalysed processes described herein. R, R^1, R^2, R^3, R^4, generic substituent. cat., catalyst.

[1]Department of Chemistry, University of Michigan, Willard Henry Dow Laboratory, 930 North University Avenue, Ann Arbor, Michigan 48109, USA.

© 2016 Macmillan Publishers Limited. All rights reserved

rely on harsh reaction conditions or the use of stoichiometric metal reagents. The earliest reported examples for carbonyl–olefin metathesis reactions involve a two-step sequence including a photoinduced [2+2] cycloaddition, followed by a subsequent acidic and/or thermal fragmentation step to induce cycloreversion to the corresponding metathesis products (Fig. 1c)[10–12]. The success of this photolysis–pyrolysis series is often impeded by harsh reaction conditions, which are prone to result in competing polymerization or regeneration of starting carbonyl–olefin pairs. Importantly, stoichiometric amounts of catalysts originally developed for olefin metathesis (Fig. 1b) were found to be able to mediate carbonyl–olefin metathesis reactions. In this case, the cycloreversion step results in the formation of a kinetically inert metal–oxo complex, precluding regeneration of the metal alkylidene and prohibiting catalyst turnover (Fig. 1c)[13].

Alternatively, titanium-based Tebbe and Petasis reagents have been used in excess to induce initial olefin metathesis and subsequent carbonyl olefination reactions, resulting in the formation of carbonyl–olefin metathesis products[14,15]. In addition to this titanium alkylidene approach, Brønsted and Lewis acids are also reported to effect carbonyl–olefin metathesis reactions of simple acetophenone and benzaldehyde derivatives, although limited in scope[16–19]. An important contribution in carbonyl–olefin metathesis reactions was reported recently, wherein an organocatalysed (3+2)/retro-(3+2) cycloaddition process allowed the ring-opening metathesis of cyclopropenes to give γ,δ-unsaturated aldehydes, which constitute formal carbonyl–olefin metathesis products (Fig. 1c)[20]. Although a remarkable advance, only highly strained cyclopropenes are applicable as olefinic components. As such, a general catalytic protocol has yet to be developed. Despite the development of the methods described earlier, the only protocol for carbonyl–olefin metathesis that has been applied to the synthesis of biologically active natural products and pharmaceutically relevant compounds[21,22] is the stoichiometric metal-mediated approach (Fig. 1c), further illustrating the need for a general catalytic system. We describe a Lewis-acid-catalysed carbonyl–olefin ring-closing metathesis reaction for the direct synthesis of functionalized carbocycles, representing a major advance in carbonyl–olefin metathesis.

At the outset of our investigations, we envisioned an efficient design principle for catalytic carbonyl–olefin metathesis based on the *in situ* formation of oxetanes as reactive intermediates (**II**) (Fig. 1c). This approach fundamentally differs from stoichiometric carbonyl–olefin metathesis protocols proceeding via intermediate metalla-oxetanes using metal-alkylidene reagents. While these result in the formation of kinetically inert metal–oxo complexes upon retro-cycloaddition that ultimately prohibit a catalytic process, the corresponding oxetane intermediates (**II**) are able to undergo direct fragmentation to the desired carbonyl–olefin metathesis products, thus enabling catalyst turnover. This conceptually simple design principle for catalytic carbonyl–olefin metathesis relies on several fundamental requirements to achieve a viable synthetic strategy. In particular, a potent catalyst needs to be identified that has the capability of efficiently promoting both the [2+2] cycloaddition as well as the subsequent [2+2] cycloreversion step. Additionally, potential competing side reactions upon [2+2] cycloreversion, such as polymerization and regeneration of starting materials, need to be reduced. Similarly, the control of chemoselectivity is challenging, as unsaturated ketone substrates are prone to undergo different modes of reactivity, including carbonyl-ene, Prins and alkylation reactions of enolate intermediates[23–26]. Initially, we hypothesized that metal-derived Lewis acids could overcome these constraints and function as effective catalysts to activate the carbonyl for both initial [2+2] cycloaddition as well as subsequent fragmentation of the resulting oxetane to yield the desired carbonyl–olefin metathesis products (**III**; Fig. 1c). However, only isolated literature reports exist for Lewis-acid-mediated oxetane formation using either equimolar amounts of titanium(IV) or tin(IV)[25,27]. In comparison, Lewis-acid-promoted oxetane fragmentation was found to be more prominent,

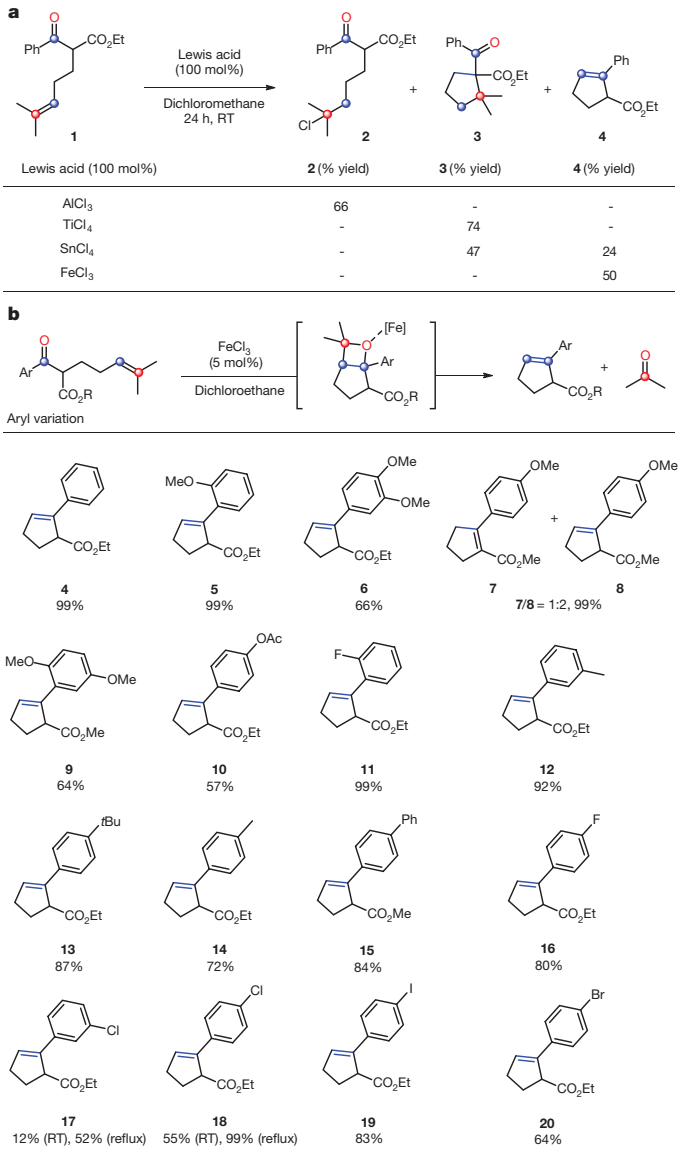


Figure 2 | Initial evaluation of the catalytic carbonyl–olefin metathesis reaction. a, Stoichiometric Lewis acid evaluation. All reactions were performed using 0.18 mmol β-ketoester **1**, 0.18 mmol Lewis acid I dichloromethane (0.1 M) at room temperature (RT) for 24 h. **b**, Evaluation of electronically differentiated arene substituents in the catalytic carbonyl–olefin metathesis reaction using iron(III) chloride as an Earth-abundant Lewis acid catalyst. Conditions: ketone (1.0 equiv.), $FeCl_3$ (5 mol%) in dichloroethane (0.01 M), room temperature, 1–12 h. **17** and **18** were run at reflux in dichloroethane. For **17** and **18**, yields at room temperature were determined by ^{1}H-NMR analysis using naphthalene as internal standard.

examples of which rely on boron trifluoride diethyl etherate, aluminium, and Brønsted acid reagents[10,27,28].

To identify a single catalyst for *in situ* oxetane formation and fragmentation, we initially investigated the reaction of β-ketoesters bearing a pendant isoprenyl group with various Lewis acids (Fig. 2). When $AlCl_3$ was used equimolar to β-ketoester **1**, the formation of the corresponding tertiary chloride **2** was observed, while stoichiometric $TiCl_4$ resulted in the tertiary alkylation product **3** in 74% yield (Fig. 2a). Upon reaction with equimolar amounts of $SnCl_4$, two new products were isolated in 47% and 24% yield, respectively. The major product of this transformation was identified as the tertiary alkylation product **3**, while the minor product was identified as the desired carbonyl–olefin metathesis product **4**, thereby providing proof of concept that a single Lewis acid is capable of promoting both intermediate oxetane

© 2016 Macmillan Publishers Limited. All rights reserved

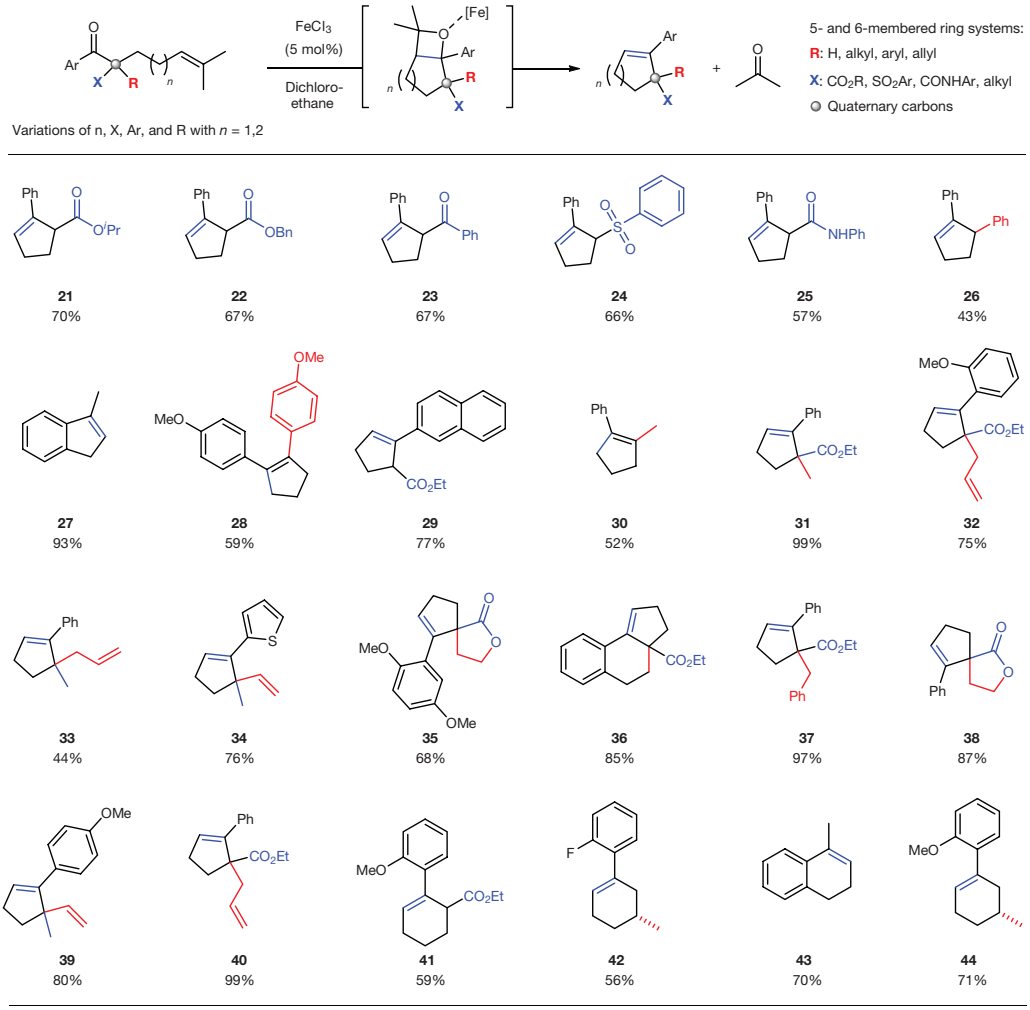


Figure 3 | Scope of the iron(III)-catalysed carbonyl–olefin metathesis reaction. Conditions: ketone (1.0 equiv.), $FeCl_3$ (5 mol%) in dichloroethane (0.01 M), room temperature, 1–12 h. **25** was stoichiometric in $FeCl_3$. **41** and **42** were run at 60 °C using $FeCl_3$ (5 mol%), 12 h. **30** was obtained as a single alkene isomer in 52% yield using $Fe(OTf)_3$ (5 mol%) at reflux and as a 1.2:1 mixture of alkene isomers in 46% yield using $FeCl_3$ (5 mol%) at room temperature. The yield for both reactions was determined by 1H-NMR analysis using naphthalene as internal standard. **28** and **44** were isolated as a mixture of alkene isomers. See Supplementary Information for details. The tetrasubstituted alkenes **28** and **30** represent the thermodynamically more stable products. For **27**, yield was determined by gas chromatography analysis using dodecane as internal standard. For **43**, yield was determined by ^{1}H-NMR analysis using naphthalene as internal standard.

formation and subsequent fragmentation, albeit in stoichiometric amounts. No reaction was observed with substoichiometric amounts of protic acids and of a variety of Lewis acids, including $FeCl_2$, $ZnCl_2$ and $SnCl_2$ (Supplementary Table 2). Succeeding investigations identified iron(III)-derived Lewis acids as superior reagents capable of selectively forming the desired carbonyl–olefin metathesis products. Ultimately, we identified 5 mol% $FeCl_3$ in dichloroethane as the optimal set of reaction conditions, resulting in quantitative formation of the desired carbonyl–olefin metathesis product (Supplementary Tables 3 and 4).

The optimized reaction conditions proved efficient for various substrates (Fig. 2b). Aromatic β-ketoesters bearing strongly electron-donating substituents on the arene subunit were converted to the corresponding cyclopentene metathesis products (**5**–**9**; Fig. 2b) in good yields. In particular, the *ortho*-methoxy cyclopentene **5** was formed quantitatively, while increasing the number of methoxy substituents led to diminished yields, with 3,4- and 2,5-dimethoxy arenes resulting in the formation of **6** and **9** in 66% and 64% yield, respectively. A mixture of *para*-substituted alkene isomers **7** and **8** was obtained in quantitative yield when the corresponding *para*-methoxy β-ketoester was subjected to otherwise identical reaction conditions.

In contrast, the *para*-acetylated cyclopentene **10** was obtained as a single product in 57% yield. β-Ketoesters bearing different substituents on the arene moiety (that is, alkyl, aryl and unsubstituted) were converted in higher yields of up to 99% to form cyclopentenes **11**–**16** (Fig. 2b). Electron-deficient chloro-, bromo- and iodoarenes resulted in the formation of the corresponding carbonyl–olefin metathesis products **17**–**20** (Fig. 2) in high yields. Taken together, these data demonstrate tolerance of various electronically differentiated arenes.

We next sought to expand the scope of the iron(III)-catalysed carbonyl–olefin metathesis reaction by investigating different β-substitution, heteroatom incorporation, and varied structural motifs (Fig. 3). First, we focused on substrates differing in their β-substitution pattern and found isopropyl and benzyl esters to be tolerated, resulting in the formation of **21** and **22** in 70% and 67% yield, respectively. Notably, ketones, sulfonyl and amide functional groups were also compatible (**23**–**25**; Fig. 3); however, stoichiometric iron was required to induce full conversion in the case of the amide. Importantly, bis-phenyl cyclopentene **26** and indene **27**, which lack β-Lewis basic substituents, are efficiently generated from the metathesis reaction. Interestingly, tetrasubstituted olefins **28** and **30** were obtained under the optimal reaction conditions as the thermodynamically stable products. Naphthyl-substituted **29** was also obtained under the optimized reaction conditions. Finally, we explored the generation of

© 2016 Macmillan Publishers Limited. All rights reserved

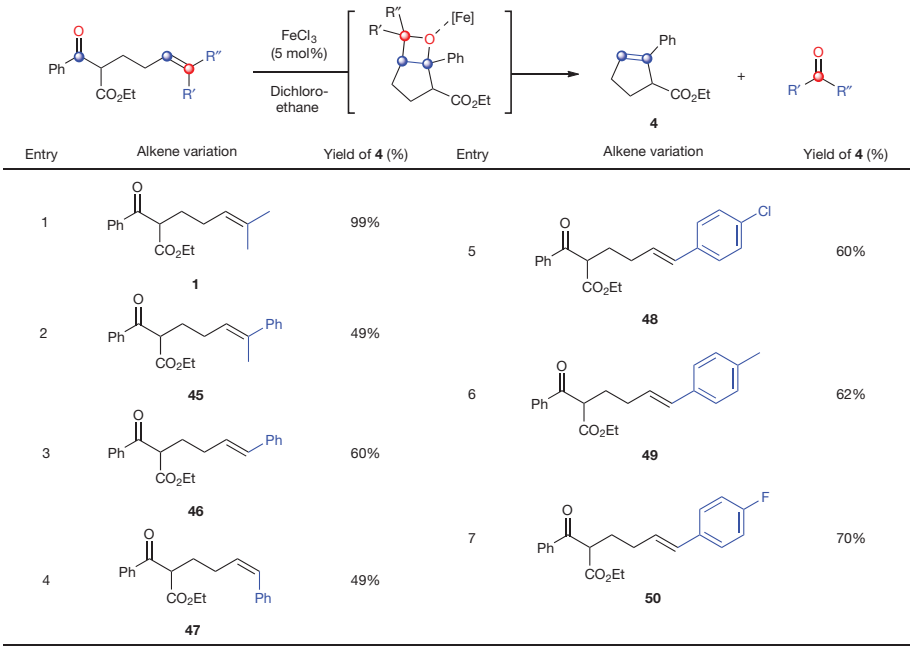


Figure 4 | Alkene evaluation in the catalytic carbonyl–olefin metathesis reaction. Conditions: ketone (1.0 equiv.), $FeCl_3$ (5 mol%) in dichloroethane (0.01 M) at room temperature. See Supplementary Information for additional experimental details. Terminal alkenes and tetrasubstituted alkenes were found to be unreactive under the optimized reaction conditions. **47**–**49** were run at reflux in dichloroethane.

cyclopentenes bearing quaternary carbon centres. Quaternary alkyl, allyl, vinyl and benzyl groups led to the formation of the corresponding cyclopentenes **31**–**40** (Fig. 3) in up to 99% yield. Additionally, when conducted on a 1.0 g scale, cyclopentene **40** was isolated in 98% yield, indicating the amenability of this protocol to being used on a larger scale. Moreover, structurally complex tricyclic and spirocyclic motifs were obtained in excellent yields in a single synthetic transformation (**35**, **36** and **38**; Fig. 3). Importantly, substrates bearing double alkyl substitution in the α-position proved effective in producing the desired metathesis adducts, resulting in the formation of cyclopentene **33** in 44%, vinyl-substituted thiophene **34** in 76%, and vinyl-anisole **39** in 80% yield. Notably, the scope of the iron(III)-catalysed carbonyl–olefin metathesis was successfully extended beyond the initial substrate scaffold to tolerate the sterically demanding bis-substitution in the α-position as well as chain elongation in the isoprenyl subunit. This set of compounds indicates that a β-Lewis basic substituent is not necessary to bind the catalyst to facilitate carbonyl activation. Furthermore, the success of substrates bearing a quaternary centre in the α-position demonstrates that the formation of a metal enolate, a process often observed upon the combination of Lewis acids and 1,3-dicarbonyls, is not necessary to form the metathesis product. Extending the olefin chain length by one carbon unit results in the formation of 6-membered carbocycles **41**–**44** in up to 71% yield.

To gain further insight into the controlling features of this process, we next turned our attention to the pendant olefin moiety. As presented (see earlier), isoprenyl-derived alkenes, such as **1** (Fig. 2a) provide efficient conversion to the metathesis product. This aspect led us to initially consider the possibility of the formation of a tertiary carbocation as a potential intermediate for catalysis. To examine this supposition, we explored the effect of varying olefin substitution (**45**–**50**; Fig. 4), which resulted in the formation of the carbonyl–olefin metathesis products in all cases in good to excellent yields. However, substrates with the potential of benzylic stabilization of an intermediate carbocation resulted in an overall decrease in yield compared to **1** (**45**–**50**; Fig. 4). These data suggest that increased stabilization of a carbocation from a simple tertiary centre can hinder product formation. Alternatively, the bulkier arene substituents present in olefins **45**–**50** compared to **1** could increase the energy of one or more of the transition states necessary for product formation.

Collectively, our examination of the substrate scope has provided initial insight into the mechanism of iron(III)-chloride-catalysed carbonyl–olefin metathesis. The reaction tolerates various aromatic substituents, while sterically encumbered olefin moieties inhibit product formation. Moreover, metal enolate formation is not required to yield the desired product, since compounds bearing quaternary carbon centres in the α-position are viable substrates for this transformation. Additionally, the formation of acetone as the corresponding carbonyl by-product was observed upon ^{1}H-NMR analysis of the reaction with olefin **1** (using $Sc(OTf)_3$ as a catalyst), consistent with the proposed metathesis process. On the basis of these initial findings, we propose two possible pathways of product formation that proceed via oxetane **E**. In a mechanism relying on carbocation intermediates, substrate **A** (Fig. 5) initially coordinates to iron(III) to form adduct **B**. This activated carbonyl can subsequently undergo nucleophilic attack by the pendant olefin on either face, resulting in the formation of carbocycles **C** and **D**. The *cis*-configuration of carbocycle **C** allows for intramolecular trapping of the carbocation to form oxetane **E**. Importantly, this cyclization is not possible for carbocycle **D**. The high yield of product implies that formation of carbocycle **D** is reversible and that it is possible for carbocycles **C** and **D** to interconvert through **B**. Oxetane **E** can undergo subsequent iron(III)-mediated ring opening to form intermediate **F**, containing a benzylic carbocation. This species can then eliminate carbonyl **H** and form the carbonyl–olefin metathesis product **G**. In an alternative mechanism, the carbonyl–olefin metathesis reaction can proceed in a concerted fashion (Fig. 5), which initially also relies on coordination of substrate **A** to iron(III) chloride to form Lewis acid adduct **B**. However, adduct **B** then reacts in a concerted [2+2]-cycloaddition to form oxetane intermediate **E**, which upon concerted [2+2]-cycloreversion results in the formation of carbonyl–olefin metathesis product **G** and carbonyl by-product **H** (Fig. 5).

Our subsequent efforts focused on developing mechanistic probe molecules to trap potential carbocation intermediates **C** and **F** upon reaction with suitable nucleophiles (Fig. 5). When substrate **1** was reacted under the optimized reaction conditions with equimolar amounts of various nucleophiles (methanol, isopropanol, acetic acid, benzoic acid), no products resulting from pathways proceeding via intermediate carbocations were observed (**51** and **52**; Fig. 5). In all

© 2016 Macmillan Publishers Limited. All rights reserved

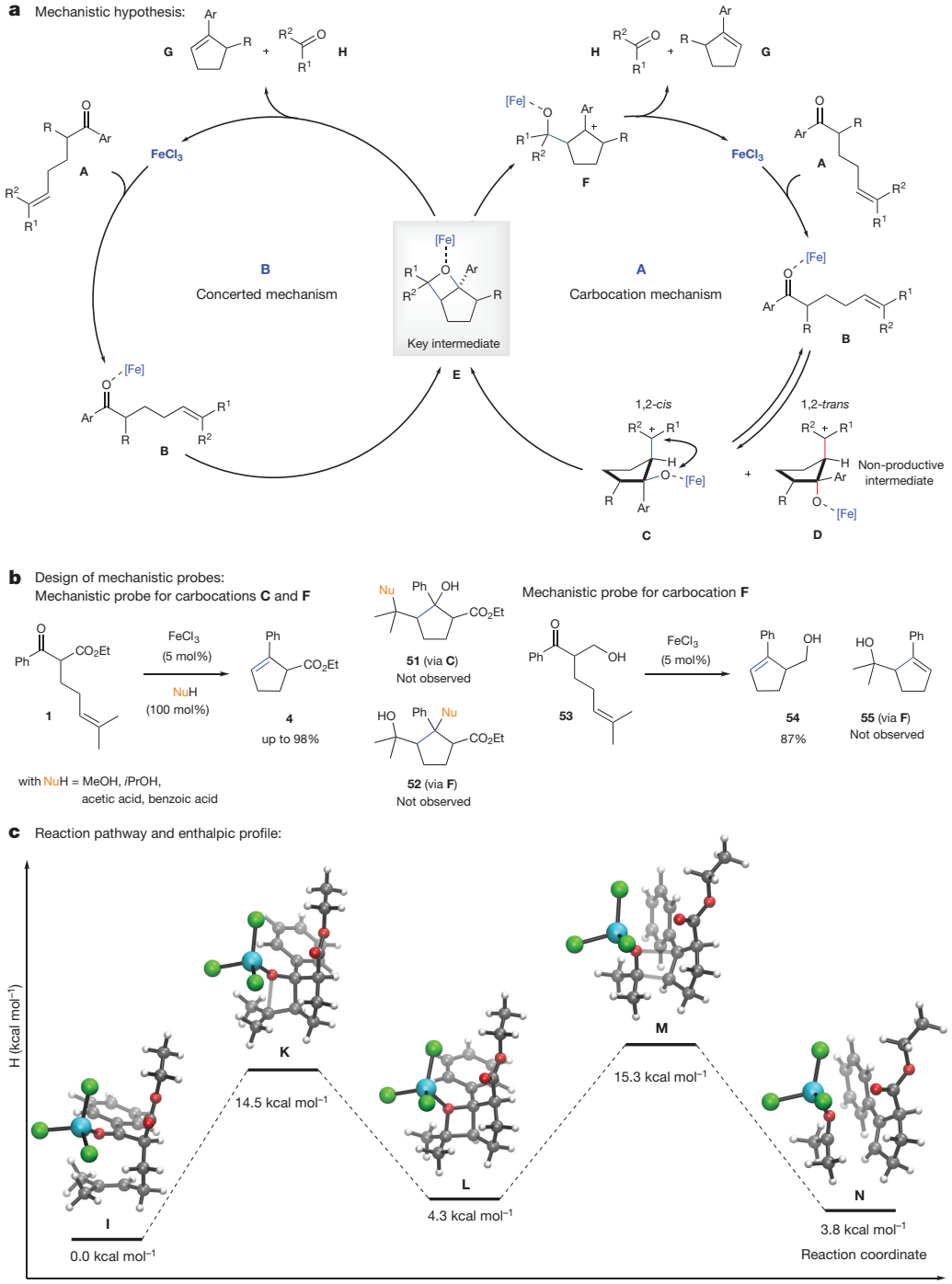


Figure 5 | Mechanistic hypothesis for the iron(III)-chloride-catalysed carbonyl–olefin metathesis reaction. a, Two possible pathways of product formation relying on either a concerted or a carbocation mechanism. **b**, Design of mechanistic probe molecules to trap potential carbocation intermediates **C** and **F**. **c**, Enthalpic profile and reaction pathway of the computationally proposed reaction mechanism for carbonyl–olefin metathesis using iron(III) chloride. Ar, generic aryl substituent; R, R^1, R^2, generic substituent.

cases, the only product obtained was the desired carbonyl–olefin metathesis product **4** in up to 98% yield. When mechanistic probe molecule **53** bearing a pendant primary alcohol was reacted, the only product isolated in the course of this transformation was the metathesis product **54**, and no product resulting from potential carbocation intermediates was observed (**55**; Fig. 5). Together, these results point towards an iron(III)-catalysed carbonyl–olefin metathesis reaction proceeding via a concerted mechanism that does not involve carbocations as essential intermediates.

We further investigated the mechanism of the carbonyl–olefin metathesis reaction computationally through the ZStruct method[29,30]. This computational tool systematically searches a combinatorial set of plausible reaction coordinates for kinetically and thermodynamically feasible elementary reactions, using quantum chemistry to provide accurate predictions for activation barriers and energies of reactions. In the case of structure **I** (Fig. 5), ZStruct generated approximately 800 reaction coordinate combinations. Upon complete quantum chemical evaluation of these reaction paths, the method successfully identified the lowest barrier transition state (**K**) leading to oxetane intermediate (**L**). This unbiased strategy differs from standard reaction simulations that typically require individual, selected mechanistic hypotheses for each step of a proposed reaction sequence.

© 2016 Macmillan Publishers Limited. All rights reserved

The full ZStruct search located one highly favourable reaction pathway, outlined in Fig. 5c. Iron(III) chloride initially binds to the carbonyl oxygen in **I** to enable formation of an oxetane intermediate via a concerted [2+2]-cycloaddition reaction with an enthalpic barrier of 14.5 kcal mol^{-1} above the iron–substrate complex **I**. This step is concerted yet asynchronous, with the carbon–carbon bond formed to a greater extent than the carbon–oxygen bond in the transition state. The resulting iron-bound oxetane **L** was found to be 4.3 kcal mol^{-1} in enthalpy above the starting complex **I**. A subsequent elementary step with a barrier of 15.3 kcal mol^{-1} allows for ring opening to result in the formation of metathesis product **N** and acetone by-product. This reaction step was also found to be asynchronous, with the carbon–oxygen bond breaking to a greater extent than carbon–carbon bond cleavage in the transition state. In each step, the carbon–oxygen distance is elongated more than the carbon–carbon bond distance in the transition state. This asynchronicity is consistent with Lewis acid activation of the carbonyl oxygen in **I**, in which the build-up of charge on the carbon–oxygen side of the forming or breaking oxetane is compensated through charge withdrawal by iron(III) chloride.

To provide additional support for the computationally proposed mechanism, rate-limiting activation barriers for ten substrates were evaluated. Seven of these substrates (**14**, **45**–**50**) were predicted to have activation barriers suitable for productive catalysis. Subsequent experiments showed a range of productivity from 0% yield for the highest barrier substrates up to 99% yield for lower barrier cases. These yields correlate well with the computational predictions, providing considerable evidence for the concerted mechanism (Fig. 5). These results also corroborate the mechanistic probe experiments (Fig. 5), where the concerted mechanism is consistent with the inability for carbocationic intermediates to be trapped.

The catalytic carbonyl–olefin metathesis reaction described here enables direct carbon–carbon bond construction between carbonyl and olefin moieties relying on iron as an environmentally benign catalyst. This new method is operationally facile, applicable to gram-scale synthesis and provides a distinct advance towards the development of a sustainable and economical approach for carbonyl–olefin metathesis.

Received 15 July 2015; accepted 17 February 2016.
Published online 27 April 2016.

1. Hoveyda, A. H. & Zhugralin, A. R. The remarkable metal-catalysed olefin metathesis reaction. *Nature* **450,** 243–251 (2007).
2. Grubbs, R. H. & Wenzel, A. G. *Handbook of Metathesis* 2nd edn, Vols 1–3 (Wiley, 2015).
3. Katz, T. J. & McGinnis, J. The mechanism of the olefin metathesis reaction. *J. Am. Chem. Soc.* **97,** 1592–1594 (1975).
4. Grubbs, R. H., Carr, D. D., Hoppin, C. & Burk, P. L. Consideration of the mechanism of the metal catalyzed olefin metathesis reaction. *J. Am. Chem. Soc.* **98,** 3478–3483 (1976).
5. Chauvin, Y. Olefin metathesis: the early days. *Angew. Chem. Int. Ed.* **45,** 3740–3747 (2006).
6. Fürstner, A. Olefin metathesis and beyond. *Angew. Chem. Int. Ed.* **39,** 3012–3043 (2000).
7. Maryanoff, B. E. & Reitz, A. B. The Wittig olefination reaction and modifications involving phosphoryl-stabilized carbanions. Stereochemistry, mechanism, and selected synthetic aspects. *Chem. Rev.* **89,** 863–927 (1989).
8. Petasis, N. A. & Bzowej, E. I. Titanium-mediated carbonyl olefinations. 1. Methylenations of carbonyl compounds with dimethyltitanocene. *J. Am. Chem. Soc.* **112,** 6392–6394 (1990).
9. Takeda, T. & Tsubouchi, A. in *Modern Carbonyl Olefination* (ed. Takeda, T.) 151–199 (Wiley, 2003).
10. Jones, G. II, Acquadro, M. A. & Carmody, M. A. Long-chain enals via carbonyl–olefin metathesis. An application in pheromone synthesis. *J. Chem. Soc. Chem. Commun.* **6,** 206–207 (1975).
11. Pérez-Ruiz, R., Miranda, M. A., Alle, R., Meerholz, K. & Griesbeck, A. G. An efficient carbonyl-alkene metathesis of bicyclic oxetanes: photoinduced electron transfer reduction of the Paternò–Büchi adducts from 2,3-dihydrofuran and aromatic aldehydes. *Photochem. Photobiol. Sci.* **5,** 51–55 (2006).
12. Valiulin, R. A. & Kutateladze, A. G. Harvesting the strain installed by a Paternò–Büchi step in a synthetically useful way: high-yielding photoprotolytic oxametathesis in polycyclic systems. *Org. Lett.* **11,** 3886–3889 (2009).
13. Fu, G. C. & Grubbs, R. H. Synthesis of cycloalkenes via alkylidene-mediated olefin metathesis and carbonyl olefination. *J. Am. Chem. Soc.* **115,** 3800–3801 (1993).
14. Stille, J. R., Santarsiero, B. D. & Grubbs, R. H. Rearrangement of bicyclo[2.2.1] heptane ring systems by titanocene alkylidene complexes to bicyclo[3.2.0] heptane enol ethers. Total synthesis of (±)-$\Delta^{9(12)}$-capnellene. *J. Org. Chem.* **55,** 843–862 (1990).
15. Iyer, K. & Rainier, J. D. Olefinic ester and diene ring-closing metathesis using a reduced titanium alkylidene. *J. Am. Chem. Soc.* **129,** 12604–12605 (2007).
16. Soicke, A., Slavov, N., Neudörfl, J.-M. & Schmalz, H.-G. Metal-free intramolecular carbonyl–olefin metathesis of ortho-prenylaryl ketones. *Synlett* **17,** 2487–2490 (2011).
17. van Schaik, H.-P., Vijn, R.-J. & Bickelhaupt, F. Acid-catalyzed olefination of benzaldehyde. *Angew. Chem. Int. Edn Engl.* **33,** 1611–1612 (1994).
18. Bah, J., Franzén, J. & Naidu, V. R. Direct organocatalytic oxo-metathesis, a *trans*-selective carbocation-catalyzed olefination of aldehydes. *Eur. J. Org. Chem.* **2015,** 1834–1839 (2015).
19. Jossifov, C., Kalinova, R. & Demonceau, A. Carbonyl olefin metathesis. *Chim. Oggi* **26,** 85–87 (2008).
20. Griffith, A. K., Vanos, C. M. & Lambert, T. H. Organocatalytic carbonyl-olefin metathesis. *J. Am. Chem. Soc.* **134,** 18581–18584 (2012).
21. Hong, B., Li, H., Wu, J., Zhang, J. & Lei, X. Total syntheses of (−)-huperzine Q and (+)-lycopladines B and C. *Angew. Chem. Int. Ed.* **54,** 1011–1015 (2015).
22. Heller, S. T., Kiho, T., Narayan, A. R. H. & Sarpong, R. Protic-solvent-mediated cycloisomerization of quinoline and isoquinoline propargylic alcohols: syntheses of (±)-3-demethoxyerythratidinone and (±)-cocculidine. *Angew. Chem. Int. Ed.* **52,** 11129–11133 (2013).
23. Clarke, M. L. & France, M. B. The carbonyl ene reaction. *Tetrahedron* **64,** 9003–9031 (2008).
24. Miles, R. B., Davis, C. E. & Coates, R. M. Syn- and anti-selective Prins cyclizations of δ,ε-unsaturated ketones to 1,3-halohydrins with Lewis acids. *J. Org. Chem.* **71,** 1493–1501 (2006).
25. Jackson, A. C., Goldman, B. E. & Snider, B. B. Intramolecular and intermolecular Lewis acid catalyzed ene reactions using ketones as enophiles. *J. Org. Chem.* **49,** 3988–3994 (1984).
26. Reetz, M. T. Lewis acid induced α-alkylation of carbonyl compounds. *Angew. Chem. Int. Edn Engl.* **21,** 96–108 (1982).
27. Demole, E., Enggist, P. & Borer, M. C. Applications synthétiques de la cyclisation d'alcools tertiaires γ-éthyléniques en α-bromotétrahydrofurannes sous l'action du N-bromosuccinimide. II. Cyclisation du (±)-nérolidol en diméthyl-2,5-(méthyl-4-pentène-3-yl)-2-cycloheptène-4-one, tétraméthyl-3, 3, 7, 10-oxa-2-tricyclo[5.5.0.01,4]-dodécène-9, β-acoratriène, cédradiène-2,8, épi-2-α-cédrène et α-cédrène. *Helv. Chim. Acta* **54,** 1845–1864 (1971).
28. Carless, H. A. J. & Trivedi, H. S. New ring expansion reaction of 2-t-butyloxetans. *J. Chem. Soc. Chem. Commun.* 382–383 (1979).
29. Zimmerman, P. M. Automated discovery of chemically reasonable elementary reaction steps. *J. Comput. Chem.* **34,** 1385–1392 (2013).
30. Zimmerman, P. M. Reliable transition state searches integrated with the growing string method. *J. Chem. Theory Comput.* **9,** 3043–3050 (2013).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by the Petroleum Research Fund (PRF#54688-DNI1) and start-up funds provided by the University of Michigan. J.R.L. thanks the National Science Foundation for a predoctoral fellowship. P.M.Z. thanks the Office of Naval Research for support under grant N00014-14-1-0551 and the Petroleum Research Fund (PRF#54267-DNI6). We are grateful to A. Speelman, A. McQuarters and N. Lehnert at the University of Michigan for helpful discussions.

Author Contributions J.R.L., J.B.G. and C.S.S. devised the experiments, prepared the starting materials and the products. P.M.Z. conducted the theoretical investigations. J.R.L., J.B.G., P.M.Z. and C.S.S. prepared this manuscript for publication.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to C.S.S. (corinnas@umich.edu).

© 2016 Macmillan Publishers Limited. All rights reserved

LETTER

doi:10.1038/nature17423

Changing atmospheric CO_2 concentration was the primary driver of early Cenozoic climate

Eleni Anagnostou[1], Eleanor H. John[2]†, Kirsty M. Edgar[2,3]†, Gavin L. Foster[1], Andy Ridgwell[4,5], Gordon N. Inglis[6,7], Richard D. Pancost[6,7], Daniel J. Lunt[4,5] & Paul N. Pearson[2]

The Early Eocene Climate Optimum (EECO, which occurred about 51 to 53 million years ago)[1], was the warmest interval of the past 65 million years, with mean annual surface air temperature over ten degrees Celsius warmer than during the pre-industrial period[2–4]. Subsequent global cooling in the middle and late Eocene epoch, especially at high latitudes, eventually led to continental ice sheet development in Antarctica in the early Oligocene epoch (about 33.6 million years ago). However, existing estimates place atmospheric carbon dioxide (CO_2) levels during the Eocene at 500–3,000 parts per million[5–7], and in the absence of tighter constraints carbon–climate interactions over this interval remain uncertain. Here we use recent analytical and methodological developments[8–11] to generate a new high-fidelity record of CO_2 concentrations using the boron isotope ($\delta^{11}B$) composition of well preserved planktonic foraminifera from the Tanzania Drilling Project, revising previous estimates[6]. Although species-level uncertainties make absolute values difficult to constrain, CO_2 concentrations during the EECO were around 1,400 parts per million. The relative decline in CO_2 concentration through the Eocene is more robustly constrained at about fifty per cent, with a further decline into the Oligocene[12]. Provided the latitudinal dependency of sea surface temperature change for a given climate forcing in the Eocene was similar to that of the late Quaternary period[13], this CO_2 decline was sufficient to drive the well documented high- and low-latitude cooling that occurred through the Eocene[14]. Once the change in global temperature between the pre-industrial period and the Eocene caused by the action of all known slow feedbacks (apart from those associated with the carbon cycle) is removed[2–4], both the EECO and the late Eocene exhibit an equilibrium climate sensitivity relative to the pre-industrial period of 2.1 to 4.6 degrees Celsius per CO_2 doubling (66 per cent confidence), which is similar to the canonical range (1.5 to 4.5 degrees Celsius[15]), indicating that a large fraction of the warmth of the early Eocene greenhouse was driven by increased CO_2 concentrations, and that climate sensitivity was relatively constant throughout this period.

Over the past 540 million years, Earth's climate has oscillated between a globally warm 'greenhouse state' and an 'icehouse state' with substantial continental glaciation[16]. The most recent of these transitions occurred between the warmest time interval of the last 65 million years—the EECO (about 14 ± 3 °C warmer than pre-industrial times[2])—and the rapid growth of ice on Antarctica in the earliest icehouse state of the Oligocene (~33.6 Myr ago[1]). It has been suggested that variations in the concentration of the greenhouse gas CO_2 were responsible for both the overall warmth of the Eocene and the subsequent cooling[17]. Recent studies have documented the importance of CO_2 decline for the final step into the icehouse across the Eocene–Oligocene transition[12,18]. Despite this, the few available CO_2 reconstructions vary markedly between different proxy systems, obscuring relationships with the global cooling trend[1,5,19,20] and therefore preventing a robust test of this hypothesis (Fig. 1). Furthermore, fully coupled climate models require CO_2 concentrations ranging from 2× to 16× the pre-industrial value (560–4,480 parts per million (p.p.m.); see, for example, ref. 21) in order to replicate reconstructed

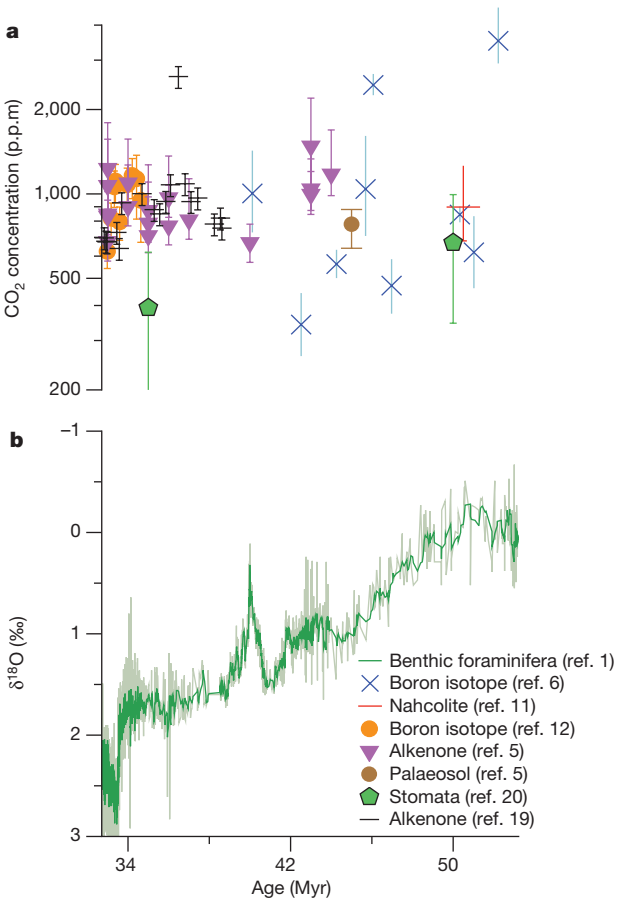


Figure 1 | Currently available Eocene atmospheric CO_2 records and benthic foraminiferal $\delta^{18}O$ values. a, CO_2 reconstructions[5,6,11,12,19,20] with the 2 s.d. uncertainties. **b**, Benthic foraminiferal $\delta^{18}O$ (ref. 1). The green line shows the five-point moving average of the grey line, with data age modified to account for the revised timing of the MECO[30]. Lower benthic $\delta^{18}O$ values denote warmer bottom-water temperatures (or less ice volume, which is assumed to be minimal in the Eocene). Error bars on $\delta^{18}O$ are smaller than the symbols (Methods).

[1]Ocean and Earth Science, National Oceanography Centre Southampton, University of Southampton Waterfront Campus, Southampton SO14 3ZH, UK. [2]School of Earth and Ocean Sciences, Cardiff University, Park Place, Cardiff CF10 3AT, UK. [3]School of Earth Sciences, Bristol University, Bristol BS8 1RJ, UK. [4]School of Geographical Sciences, Bristol University, Bristol BS8 1SS, UK. [5]Department of Earth Sciences, University of California, Riverside, California 92521, USA. [6]Organic Geochemistry Unit, School of Chemistry, University of Bristol, Cantock's Close, Bristol BS8 1TS, UK. [7]Cabot Institute, University of Bristol, Bristol BS8 1UJ, UK. †Present addresses: School of Geography, Earth Sciences and Environment, University of the South Pacific, Suva, Fiji (E.H.J.); School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham B15 2TT, UK (K.M.E.).

© 2016 Macmillan Publishers Limited. All rights reserved

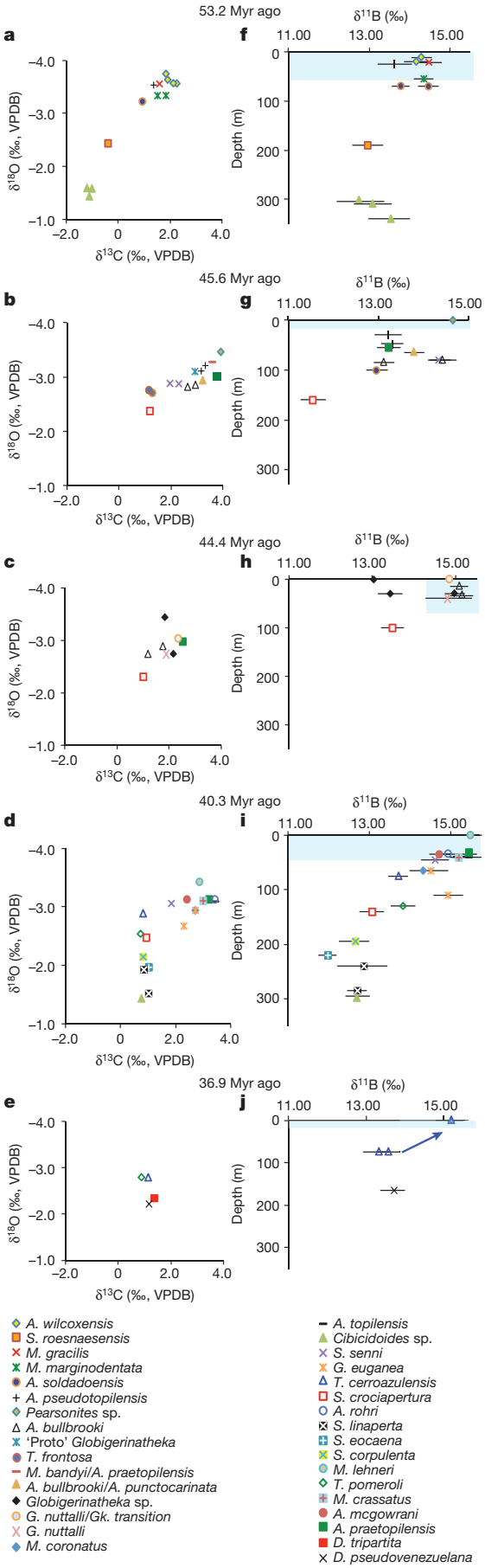


Figure 2 | Eocene planktonic foraminiferal multi-species stable isotope arrays. a–e, Foraminiferal $\delta^{13}C$ versus $\delta^{18}O$ for each time slice. **f–j**, Foraminiferal $\delta^{11}B$ values versus $\delta^{18}O$ inferred water depths for each time slice. **h** shows the anomalous behaviour of *Globigerinatheka* spp. (black diamonds) in both $\delta^{13}C$ and $\delta^{11}B$ (see Methods). The blue arrow in **j** shows the correction of *T. cerroazulensis* $\delta^{11}B$ to shallower water values (Methods). Shaded areas in **f–j** surround the species used in Fig. 3a. Errors on $\delta^{13}C$ and $\delta^{18}O$ are smaller than symbols. Error bars on $\delta^{11}B$ are 2 s.d. based on long-term precision (Methods). VPDB, the Vienna Pee-Dee belemnite standard. For a full species list see Supplementary Table 1.

early Eocene sea surface temperatures (SSTs) and mean annual air temperatures (MATs). Clearly, to use past warm climate states like the early Eocene to evaluate models and our understanding of the climate system, new and more accurate CO_2 reconstructions are needed.

The use of boron isotopes ($\delta^{11}B$) in marine carbonates is a well understood proxy of seawater pH, allowing for high-fidelity reconstructions of atmospheric CO_2 in locations where surface seawater is in near-equilibrium with the atmosphere regarding CO_2 (see refs 9 and 22). Recently, the isotopic fractionation factor of boron in seawater has been experimentally determined[10], our understanding of the evolution of the ocean carbonate system through time has improved[23,24], and it has become apparent through inter-laboratory comparison studies[8] that a potential bias relating to the foraminiferal B/Ca ratio may exist in earlier studies, suggesting that published Eocene $\delta^{11}B$–CO_2 estimates[6,7] need to be revisited. Here we resolve these issues by measuring the $\delta^{11}B$ of well preserved Eocene planktonic foraminifera from the Kilwa group in Tanzania using multi collector inductively coupled plasma mass spectrometry (MC-ICPMS)[9] and a greater array of species, to resolve the potential for species-specific and depth-habitat-related biases in $\delta^{11}B$ more clearly.

The approach we follow to determine atmospheric CO_2 from foraminiferal calcium carbonate $\delta^{11}B$ ($\delta^{11}B_c$) is detailed in the Methods. We analysed foraminifera from five discrete time slices between 36.9 and 53.2 Myr ago recovered by the Tanzania Drilling Project (Extended Data Fig. 1), and we quantified the oxygen isotopic composition ($\delta^{18}O$) of up to 17 different foraminifera species occupying a range of depth habitats to derive the calcification temperature and hence the relative habitat depth of the taxa[25]. In each case we find a decrease of $\delta^{11}B_c$ with increasing depth, consistent with modern ocean $\delta^{11}B_{borate}$ profiles (Fig. 2) and a clustering of the warmest species with a similar $\delta^{11}B_c$.

By applying species-specific $\delta^{11}B$–pH calibrations, $\delta^{11}B_c$ measurements can be converted to pH reconstructions with depth. However, we lack such calibrations for these extinct taxa. The $\delta^{11}B$ vital effects that are a result of foraminiferal physiology (Methods) appear to be reduced in Eocene foraminifera compared to modern taxa for both shallow symbiont-bearing and deeper non-symbiotic planktonic foraminifera. We therefore calculate pH from $\delta^{11}B_c$ using two different approaches to bracket the probable magnitude of $\delta^{11}B$–pH vital effects: (1) we assume that $\delta^{11}B_c = \delta^{11}B_{borate}$ (hereafter the 'borate' calibration), and (2) we assume that the shallower foraminifera in each time slice has the $\delta^{11}B$–pH calibration of the modern mixed-layer-dwelling foraminifera *Trilobatus sacculifer*[22].

The boron isotopic composition of seawater ($\delta^{11}B_{sw}$) is also required to calculate pH, constrained here by two methods, both based on the observation that the reconstructed pH gradient from shallow to deep for a given $\delta^{11}B_{borate}$ gradient in the water column is a function of $\delta^{11}B_{sw}$ because of the nonlinearity of the $\delta^{11}B$–pH proxy (Extended Data Fig. 2, Methods). If we assume $\delta^{11}B_c = \delta^{11}B_{borate}$ we calculate $\delta^{11}B_{sw}$ to be in the range 38.2–38.7‰. This is slightly increased to 38.6–38.9‰ if we use the $\delta^{11}B$–pH calibration of *T. sacculifer* (Methods).

Given the strong positive relationship between $\delta^{11}B$ and pH, it is evident from our $\delta^{11}B_c$ data alone that surface water pH increased through the Eocene (Fig. 3a). However, having determined both $\delta^{11}B_{sw}$ and calcification temperature, we can generate quantitative estimates of upper-ocean pH, and here we propagated a conservative estimate of

© 2016 Macmillan Publishers Limited. All rights reserved

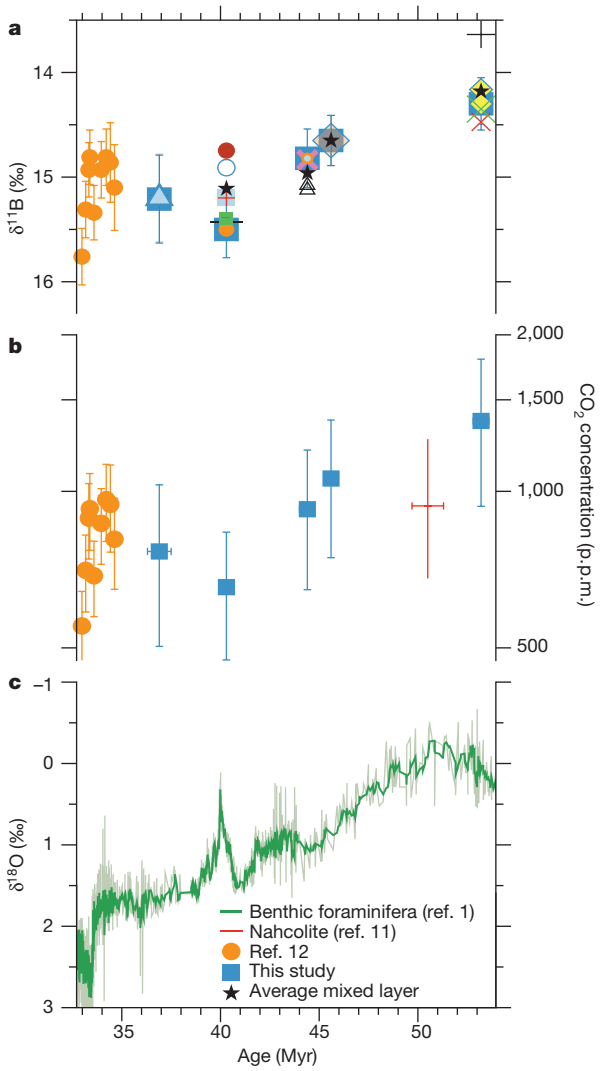


Figure 3 | New atmospheric CO_2 reconstructions from shallow planktonic foraminiferal $\delta^{11}B$. a, $\delta^{11}B$ of shallower foraminifera (symbols as in blue-shaded areas of Fig. 2) and asterisks represent the average of those for each time slice. Blue squares and orange circles represent the warmest species used in **b**. Ref. 12 data are offset (Supplementary Table 2, Methods). **b**, Atmospheric CO_2 assuming $\delta^{11}B_c = \delta^{11}B_{borate}$ using the warmest species of each time slice. The red line is based on ref. 11. **c**, Benthic foraminiferal $\delta^{18}O$ (ref. 1) (as in Fig. 1b). Age error bars (**b**) are based on the ages of the nearest datums, $\delta^{11}B$ errors are as in Fig. 2, and the CO_2 uncertainty is based on the 95% confidence of Monte Carlo error propagation (Methods).

the relevant uncertainties using a Monte Carlo approach (Methods). Based on these upper-water-column reconstructions and the borate (or *T. sacculifer)* calibration, we estimate that surface water pH increased from 7.66 ± 0.06 (or 7.59 ± 0.06) at 53.2 Myr ago, to 7.80 ± 0.07 (or 7.76 ± 0.06) by 36.9 Myr ago. By extending our knowledge of species depth habitat and $\delta^{11}B_{sw}$ to the $\delta^{11}B_c$ data set of ref. 12 (Methods), it is apparent that surface-water pH continued to increase in the early Oligocene, reaching 7.88 ± 0.05 (or 7.84 ± 0.05) by ~33 Myr ago. Notably, our Eocene surface-water pH estimates are lower than any others reported in the Cenozoic era, including during the warmth of the middle Miocene and Pliocene: minimum pH values of 7.83 (ref. 22) and 7.99 (ref. 26), respectively.

Ocean pH, like seawater CO_2—that is, $[CO_2]_{aq}$—is largely determined by the ratio of total alkalinity to total dissolved inorganic carbon. Hence, to first order, our pH estimates also indicate that $[CO_2]_{aq}$ declined through the Eocene. Given the oligotrophic setting of our site, this further implies that atmospheric CO_2 declined through the study interval. However, quantitative constraints on early Cenozoic CO_2 evolution require a second carbonate system parameter. Here we assume that the surface seawater saturation state for calcite (Ω_{calc}) at our site remained between 5.5 and 7.5, a conservative range that encompasses geochemical model estimates[27] (Methods; Extended Data Fig. 3). Uncertainty was again fully propagated (details in Methods).

As expected given the raw $\delta^{11}B_c$ values and calculated pH, we find that CO_2 decreased (Fig. 3b) for the borate calibration (or the *T. sacculifer* calibration in Extended Data Fig. 4) from $1,400 \pm 470$ p.p.m. (or $1,900 \pm 600$) p.p.m. during the EECO to $\leq$770 p.p.m. (or 920 p.p.m.) from 40.3 Myr ago onwards, reaching a minimum of 550 ± 190 p.p.m. (or 670 ± 210 p.p.m.) in the early Oligocene. Our new data confirm that CO_2 was most probably $\geq$1,000 p.p.m. (at a 95% confidence interval) during the EECO, consistent with the recently revised estimates using the sodium carbonate mineral nahcolite[11]. Although our younger time slices are in good agreement with alkenone-based CO_2 reconstructions (Extended Data Fig. 4), our early Eocene CO_2 reconstructions are clearly unprecedented in the Cenozoic. At 5–7× the pre-industrial value, our reconstructed EECO CO_2 value can provide tighter constraints on models than those that have been previously available. Critically, our CO_2 reconstructions are lower than is required by a number of coupled ocean–atmosphere climate models to achieve a good match with Eocene temperature data[21].

Using our new CO_2 record, the possible role of CO_2 in driving Eocene cooling can be quantitatively examined. The relative change in climate forcing attributable to CO_2 change ($\Delta F_{CO_2\text{-vs-EECO}}$) through the Eocene can be estimated[28] as:

$$\Delta F_{CO_2\text{-vs-EECO}} = 5.32\ln(C_t / C_{EECO}) + 0.39[\ln(C_t / C_{EECO})]^2 \quad (1)$$

where C_{EECO} is the atmospheric CO_2 concentration at the EECO, and C_t refers to the CO_2 reconstruction at a particular time in the Eocene. The CO_2 forcing based on our reconstructions for the Eocene is well correlated with the contemporaneous proxy-derived sea surface cooling, albeit with considerably larger cooling at high latitudes (Fig 4a, Extended Data Fig. 5a, $R^2 = 0.8$ at low latitudes and 0.9 at high latitudes).

The apparent latitudinal dependency of SST change for a given change in radiative forcing has recently been determined for the last 520,000 years (that is, 520 kyr; ref. 13), using CO_2, ice-sheet albedo (which are slow feedbacks), dust and mean annual insolation as forcing terms[29]. Assuming that this latitudinal dependency for a given forcing (Fig. 4b and Methods)[13] is also appropriate for the Eocene, we can calculate the expected SST change as a result of the climate forcing, ΔF_{CO_2} (see Equation (1)), for each target time slice in our study, for high latitudes (60° N and 60° S; blue lines on Fig. 4b) and low latitudes (30° N to 30° S, red box on Fig. 4b; see Methods) and compare these to a recent SST compilation[14]. The calculated SST change from our CO_2 reconstructions is within error of the observed, proxy-derived SST evolution of both high and low latitudes through the Eocene (Fig. 4c, Extended Data Fig. 5c) regardless of which $\delta^{11}B$–pH calibration is used. This supports CO_2 as the principal driver of long-term Eocene cooling, although feedback processes involved in causing the observed polar amplification in Fig. 4a and Extended Data Fig. 5a must also play a vital part in determining its pattern.

Our CO_2 reconstructions also allow us to examine the sensitivity of the Earth's temperature to changes in radiative forcing. This is often discussed in terms of Equilibrium Climate Sensitivity (ECS): the global mean surface temperature change (in degrees Celsius) for a doubling of CO_2 after all those feedbacks that operate on the timescale of about a century have had time to act (the 'fast' feedbacks[29]). The ECS can be calculated for the Eocene using:

$$ECS = \Delta T / \Delta F_{CO_2\text{-vs-PI}} \times 3.87 \quad (2)$$

© 2016 Macmillan Publishers Limited. All rights reserved

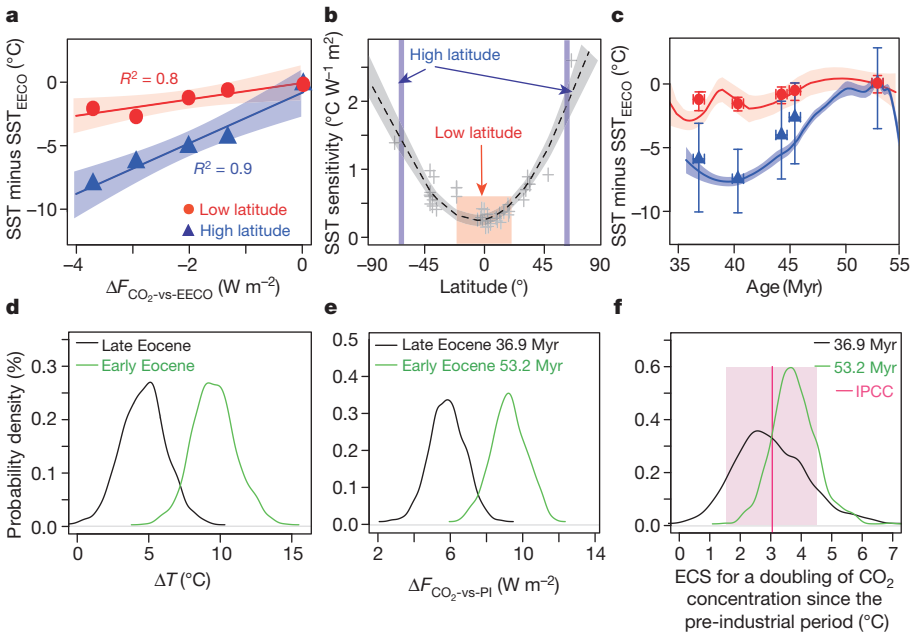


Figure 4 | CO_2 as a driver of latitudinal cooling in the Eocene, and ECS analyses of the EECO and late Eocene time slices. a, Evolving relationship between SST[14] for high and low latitudes and the CO_2 forcing of each of our time slices relative to the EECO, with linear regression fits and coefficients of determination (R^2). **b**, Apparent latitudinal SST sensitivity for the last 520 kyr (ref. 13). The dashed line is the second-order polynomial through the SST sensitivity data (grey crosses) of ref. 13. Red shading shows the SST sensitivity estimates averaged as a low-latitude mean, and blue lines show the high-latitude mean. **c**, Reconstructed (lines) and estimated (symbols) SST relative to 53.2 Myr ago. Symbols are estimated using the respective CO_2 reconstructions and the average low-latitude (red circles) and high-latitude (blue triangles) SST sensitivities of **b**. Bold lines show the reconstructed long-term mean SST estimates using the TEX_{86} proxy[14] at high latitudes (blue >55°) and low latitudes (red <30°) relative to the SST ~53 Myr ago. Error bars represent full propagation of errors at 95% confidence. **d**, Range in mean surface temperature change ΔT for early and late Eocene, corrected to remove the changes due to slow feedbacks[2,4,21,29]. **e**, Forcing compared to the pre-industrial period calculated using our CO_2 reconstructions for the time slices 53.2 Myr ago (early Eocene) and 36.9 Myr ago (late Eocene). **f**, Probability density functions of ECS for the early and late Eocene compared to IPCC estimates. Shaded bands around lines throughout show 95% confidence intervals.

where ΔT is the global mean surface temperature change compared to the pre-industrial period, corrected for the action of those feedbacks that operate on timescales longer than a century (the 'slow' feedbacks[29]) (Fig. 4d, Extended Data Fig. 5d); $\Delta F_{CO_2\text{-vs-PI}}$ is calculated as in Equation (1), replacing C_{EECO} with pre-industrial CO_2 (278 p.p.m.) because we are interested in forcing change relative to the pre-industrial period (Fig. 4e, Extended Data Fig. 5e); and 3.87 is the climate forcing of a CO_2 doubling (in units of W m^{-2}), calculated using Equation (1).

The global mean surface temperature change for the EECO is thought to be ~14 ± 3 °C warmer than the pre-industrial period, and ~5 °C warmer than the late Eocene (35 Myr ago; refs 2 and 29). Recent climate modelling studies have ascribed 4–6 °C of Eocene warmth to changes in boundary conditions (such as palaeotopography and continental configuration) and the action of the slow climate feedbacks (principally the lack of a continental ice sheet on Antarctica and vegetation change[2–4]). If atmospheric CO_2 concentration were the principal driving mechanism for the relative warmth of the EECO and late Eocene, we would expect the Eocene ECS value to fall within the range reported by the Intergovernmental Panel on Climate Change[15], based on climate modelling and observations of the (palaeo)climate system. However, if other climate feedbacks were important that we have not accounted for in our correction of the 'slow' feedbacks[2–4], or if the forcing is not due to CO_2 alone, then the calculated ECS value would be different from that reported by the IPCC. Full details of how the uncertainties in these various parameters are propagated can be found in the Methods, but from Fig. 4f and Extended Data Fig. 5f it is clear that the average ECS we calculate for the Eocene lies, for the borate (or *T. sacculifer*) approach in the range 2.6–4.6 °C (or 2.1–3.6 °C) per CO_2 doubling (66% confidence) with maximum probability for the EECO of 3.8 °C (or 3.3 °C). This range (2.1–4.6 °C per CO_2 doubling) is similar to the IPCC range (1.5–4.5 °C at 66% confidence[15]). Therefore, the 65% of early and late Eocene warmth attributed to fast feedbacks[2–4,29] is consistent with Eocene warmth being driven by higher atmospheric CO_2 concentrations. This supports the importance of atmospheric CO_2 variability in determining both the warmth of the early Eocene and the cooling that ultimately led to the establishment of the late Cenozoic icehouse.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 23 March 2015; accepted 10 February 2016.
Published online 25 April; corrected online 18 May 2016 (see full-text HTML version for details).

1. Zachos, J. C., Dickens, G. R. & Zeebe, R. E. An early Cenozoic perspective on greenhouse warming and carbon-cycle dynamics. *Nature* **451,** 279–283 (2008).
2. Caballero, R. & Huber, M. State-dependent climate sensitivity in past warm climates and its implications for future climate projections. *Proc. Natl Acad. Sci. USA* (2013).
3. Lunt, D. J. *et al.* CO_2-driven ocean circulation changes as an amplifier of Paleocene-Eocene thermal maximum hydrate destabilization. *Geology* **38,** 875–878 (2010).
4. Loptson, C. A., Lunt, D. J. & Francis, J. E. Investigating vegetation-climate feedbacks during the early Eocene. *Clim. Past* **10,** 419–436 (2014).
5. Beerling, D. J. & Royer, D. L. Convergent Cenozoic CO_2 history. *Nature Geosci.* **4,** 418–420 (2011).
6. Pearson, P. N. & Palmer, M. R. Atmospheric carbon dioxide concentrations over the past 60 million years. *Nature* **406,** 695–699 (2000).
7. Pearson, P. N. & Palmer, M. R. Middle Eocene seawater pH and atmospheric carbon dioxide concentrations. *Science* **284,** 1824–1826 (1999).
8. Foster, G. L. *et al.* Interlaboratory comparison of boron isotope analyses of boric acid, seawater and marine $CaCO_3$ by MC-ICPMS and NTIMS. *Chem. Geol.* **358,** 1–14 (2013).
9. Foster, G. L. Seawater pH, pCO_2 and [CO_3^{2-}] variations in the Caribbean Sea over the last 130 kyr: a boron isotope and B/Ca study of planktic foraminifera. *Earth Planet. Sci. Lett.* **271,** 254–266 (2008).
10. Klochko, K. *et al.* Experimental measurement of boron isotope fractionation in seawater. *Earth Planet. Sci. Lett.* **248,** 276–285 (2006).

© 2016 Macmillan Publishers Limited. All rights reserved

11. Jagniecki, E. A., Lowenstein, T. K., Jenkins, D. M. & Demicco, R. V. Eocene atmospheric CO_2 from the nahcolite proxy. *Geology* **43,** 1075–1078 (2015).
12. Pearson, P. N., Foster, G. L. & Wade, B. S. Atmospheric carbon dioxide through the Eocene–Oligocene climate transition. *Nature* **461,** 1110–1113 (2009).
13. Rohling, E. J. *et al.* Sea surface and high-latitude temperature sensitivity to radiative forcing of climate over several glacial cycles. *J. Clim.* **25,** 1635–1656 (2012).
14. Inglis, G. N. *et al.* Descent toward the Icehouse: Eocene sea surface cooling inferred from GDGT distributions. *Paleoceanography* **30,** 1000–1020 (2015).
15. Intergovernmental Panel on Climate Change (IPCC) *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Stocker, T. F. *et al.*) 1–1535 (2013).
16. Crowell, J. C. & Frakes, L. A. Phanerozoic glaciation and the causes of ice ages. *Am. J. Sci.* **268,** 193–224 (1970).
17. Raymo, M. E. & Ruddiman, W. F. Tectonic forcing of late Cenozoic climate. *Nature* **359,** 117–122 (1992).
18. Pagani, M. *et al.* The role of carbon dioxide during the onset of Antarctic glaciation. *Science* **334,** 1261–1264 (2011).
19. Zhang, Y. G. *et al.* A 40-million-year history of atmospheric CO_2. *Phil. Trans. R. Soc. A* **371,** http://dx.doi.org/10.1098/rsta.2013.0096 (2013).
20. Franks, P. J. *et al.* New constraints on atmospheric CO_2 concentration for the Phanerozoic. *Geophys. Res. Lett.* **41,** http://dx.doi.org/10.1002/2014gl060457 (2014).
21. Lunt, D. J. *et al.* Warm climates of the past—a lesson for the future? *Phil. Trans. R. Soc. A* **371,** http://dx.doi.org/10.1098/rsta.2013.0146 (2013).
22. Foster, G. L., Lear, C. H. & Rae, J. W. B. The evolution of pCO_2, ice volume and climate during the middle Miocene. *Earth Planet. Sci. Lett.* **341–344,** 243–254 (2012).
23. Hönisch, B. *et al.* The geological record of ocean acidification. *Science* **335,** 1058–1063 (2012).
24. Demicco, R. V., Lowenstein, T. K. & Hardie, L. A. Atmospheric pCO_2 since 60 Ma from records of seawater pH, calcium, and primary carbonate mineralogy. *Geology* **31,** 793–796 (2003).
25. John, E. H. *et al.* Warm ocean processes and carbon cycling in the Eocene. *Phil. Trans. R. Soc. A* **371,** http://dx.doi.org/10.1098/rsta.2013.0099 (2013).
26. Martínez-Boti, M. A. *et al.* Plio-Pleistocene climate sensitivity evaluated using high-resolution CO_2 records. *Nature* **518,** 49–54 (2015).
27. Ridgwell, A. & Schmidt, D. N. Past constraints on the vulnerability of marine calcifiers to massive carbon dioxide release. *Nature Geosci.* **3,** 196–200 (2010).
28. Byrne, B. & Goldblatt, C. Radiative forcing at high concentrations of well-mixed greenhouse gases. *Geophys. Res. Lett.* **41,** 152–160 (2014).
29. PALAEOSENS Project Members. Making sense of palaeoclimate sensitivity. *Nature* **491,** 683–691 (2012); erratum **494,** 130 (2013).
30. Bohaty, S. M., Zachos, J. C., Florindo, F. & Delaney, M. L. Coupled greenhouse warming and deep-sea acidification in the middle Eocene. *Paleoceanography* **24,** PA2207 (2009).

Supplementary Information is available in the online version of the paper.

Acknowledgements Financial support was provided by NERC grants (NE/H017356/1 and NE/I005595/1 to G.L.F. and P.N.P.) and by a NERC Post Doctoral Research Fellowship (NE/H016457/1 to K.M.E.). G.N.I. thanks the UK NERC for supporting his PhD studentship (via NE/I005595/1) and R.D.P. acknowledges the Royal Society Wolfson Research Merit Award. R.D.P. and G.N.I. also acknowledge the Advanced ERC Grant T-GRES (340923). We thank the Tanzania Petroleum Development Corporation, the Tanzania Commission for Science and Technology and the Tanzania Drilling Project field team for support. We also acknowledge A. Milton and S. Nederbraght for technical assistance, and we are grateful to M. Huber for discussions on drivers of Eocene warmth.

Author Contributions E.A. conducted all boron isotope and trace element analyses, performed calculations, and drafted the manuscript. E.H.J. and K.M.E. prepared foraminifer samples and conducted the stable isotope analysis. P.N.P. led the fieldwork, performed the taxonomy and prepared foraminifer samples. A.R. provided cGENIE model results. P.N.P. and G.L.F. designed the study and all authors discussed the results and contributed to the final text.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.A. (e.anagnostou@noc.soton.ac.uk).

© 2016 Macmillan Publishers Limited. All rights reserved

METHODS

Site information. To obtain optimal records of oxygen, carbon, and boron isotopes, we used well preserved 'glassy' carbonate microfossils from the Palaeogene of the Kilwa Group material drilled by the Tanzanian Drilling Project[31,32]. The core sites were located on the outer shelf or upper continental slope approximately 50–70 km offshore[33], at a palaeolatitude of 19° S, and bathyal palaeowater depths of >300 m (ref. 31). Seismic data, sedimentary facies, nannofossil and planktic foraminiferal assemblages as well as low concentrations of dinoflagellates and a lack of marine biomarkers indicative of high export productivity confirm that these sediments were deposited under open-ocean conditions with waters derived from the Indian Ocean subtropical gyre[34–36], supporting the assumption that $[CO_2]_{aq}$ offshore Tanzania was in (near) equilibrium with the atmosphere with respect to CO_2. This is supported by the 'cGENIE' Earth system model estimates with Eocene boundary conditions for offshore Tanzania (Extended Data Fig. 6)[37]; nevertheless, to include the uncertainty of potential disequilibrium with the atmosphere we added ±40 p.p.m. CO_2 uncertainty into our estimates of atmospheric CO_2 using $\delta^{11}B$.

Sample preparation. Approximately 3–4 mg of mono-specific foraminiferal material of a narrow size fraction (for example, ~300 individuals from the 212–250 μm sieve size fraction) were separated from ~10 cm of core material (24 cm for the time slice 53.2 Myr ago) for tandem analyses of all isotopes ($\delta^{18}O$, $\delta^{13}C$ and $\delta^{11}B$) and trace-element composition. Identification of planktonic foraminifera followed ref. 35. Each sample was gently crushed and divided into two parts: one for $\delta^{11}B$ and elemental analyses and another for $\delta^{18}O$ and $\delta^{13}C$ measurements (~100 μg, ~3% per weight). For the timeslices 53.2 Myr ago and 40.3 Myr ago, *Cibicidoides* sp. were also picked for analyses (~1 mg). Splits for elemental and $\delta^{11}B$ analyses were cleaned following established methods[38–40]. Trace-element ratios were determined as in ref. 9 and in all cases Al/Ca ratios were <120 μmol mol^{-1} (typically <100 μmol mol^{-1}), showing efficient foraminiferal cleaning[9,40].

$\delta^{18}O$–$\delta^{13}C$ analyses, and temperature and relative depth assignments. For all $\delta^{18}O$ and $\delta^{13}C$ analyses, we followed the analytical approach described elsewhere[41], and values are reported relative to the Vienna Pee Dee Belemnite (VPDB) standard, with 1 s.d. standard external precision of ±0.06‰ and ±0.07‰ respectively. All temperature reconstructions are based on the equation of ref. 42, corrected for changes in global ice volume of −0.8‰[43] and palaeolatitude of +0.83‰[44] (following ref. 45). Any further secular trends in seawater $\delta^{18}O$, regional variations in the oxygen isotope ratio of seawater, or differing carbonate ion concentrations are not known with sufficient confidence, so no additional corrections were applied. The lowest $\delta^{18}O$ value in each multi-species depth profile was used to calculate the maximum SST for each time slice with a conservative uncertainty of ±2 °C that encompasses the differences among variable carbonate $\delta^{18}O$ equations (ref. 45), and includes potential uncertainties in $\delta^{18}O_{sw}$ (refs 43 and 46).

Identification of foraminiferal depth habitats is based on $\delta^{18}O$ following refs 25 and 35. This method could carry large uncertainties in the absolute depth reconstructions, but the relative depth habitats of surface and deep taxa are clearly distinguishable from the $\delta^{13}C$ versus $\delta^{18}O$ cross plots (see, for example, refs 25, 45, 47–54), and supported by comparison of $\delta^{11}B$ versus $\delta^{18}O$ profiles (Fig. 2). Further support for our approach is given by the depth assignment of benthic foraminifera *Cibicidoides* based on $\delta^{18}O$ values at 53.2 and 40.3 Myr ago, which at 300–320 m agrees well with the site's palaeowater depth[31]. Any uncertainty in the absolute depth assignment of foraminifera as a result of the 4 °C range (±2 °C) contributes <0.1% uncertainty to the carbonate system calculations and thus is not included, although the temperature uncertainty is fully propagated into our carbonate system estimates (see below).

Boron isotope proxy and analyses. Boron isotopes in planktonic foraminifera residing close to the seawater surface have been used extensively to reconstruct past ocean pH and thus CO_2 concentrations[9,22,55–57]. External reproducibility of $\delta^{11}B_c$ analyses is calculated using the approach of ref. 40 and the relationship of ref. 41.

To reconstruct seawater pH and thus CO_2 concentrations on million-year timescales using foraminiferal $\delta^{11}B$ values, we must take some steps that include: (1) correcting for any vital effects in extinct species; (2) constraining the $\delta^{11}B_{sw}$, which, given the 10–20 Myr ago residence time for boron in the ocean is probably different from the modern value (39.61‰)[58,59]; and (3) estimating a second carbonate system parameter to infer aqueous CO_2 concentration ($[CO_2]_{aq}$) from pH. Other necessary steps include (4) reconstructing the calcification temperature and relative habitat depth of foraminifera; (5) knowing the CO_2 disequilibrium between surface waters and atmosphere at the study site; and (6) ensuring that primary $\delta^{11}B$ values are not altered by diagenesis (mitigated here by using 'glassy' foraminifera).

Vital-effects corrections associated with the boron isotope proxy. Offsets between the boron isotopic composition of ambient seawater borate ion ($\delta^{11}B_{borate}$) and that of foraminiferal calcite ($\delta^{11}B_c$) are known as vital effects. Such vital effects develop in response to respiration, calcification and photosynthesis generating a pH gradient within the micro-environment (the diffusive boundary layer) around foraminifera[60,61]. For extant species these effects can be accounted for using species-specific $\delta^{11}B$–pH calibrations (for example, refs 9, 62–64). However, all Eocene foraminifera species are extinct and thus vital effects can only be assessed indirectly. To supplement our existing understanding of modern vital effects[22,26,62,64,65] we generated core-top (Holocene) planktonic foraminiferal $\delta^{11}B$, $\delta^{18}O$ and $\delta^{13}C$ values (Extended Data Fig. 7c and Supplementary Table 1) from offshore Tanzania (Glow 15, ref. 66). Modern foraminifera from offshore Tanzania show a substantially less coherent relationship with $\delta^{18}O$ than our Eocene samples (Extended Data Fig. 7 and Fig. 2). This increased spread is at least partly due to the known large vital effects of modern foraminifera, where offsets of 0–4‰ are required for the $\delta^{11}B$ of our modern species to match the $\delta^{11}B$ of seawater borate (Extended Data Fig. 7c).

If we apply published modern symbiont-bearing planktonic foraminifera $\delta^{11}B$–pH calibrations[22,62,64] for the shallower Eocene species, and symbiont-barren $\delta^{11}B$–pH calibrations of modern planktonic foraminifera[26,65] for the Eocene deep dwellers, an unrealistic picture emerges (Extended Data Fig. 8). Specifically, there is increased scatter between the shallowest and deepest planktonic foraminifera and inversion in pH gradients with depth. Additionally, there is clear disagreement between *Cibicidoides*, that we know exhibits minimal vital effect for $\delta^{18}O$, $\delta^{11}B$ and $\delta^{13}C$ (refs 40, 67), and deeper planktonic foraminifera of similar depth (Extended Data Fig. 8), which are otherwise overlapping in $\delta^{11}B_c$ (for the 53.2-Myr-ago time slice, *S. roesnaesensis* $\delta^{11}B = 12.96 \pm 0.38$ ‰ is in agreement with *Cibicidoides*, which average 13.15 ± 0.53‰, and for the 40.3-Myr-ago time slice, the average *Subbotina* $\delta^{11}B = 12.63 \pm 0.38$‰ is within uncertainty of the *Cibicidoides* $\delta^{11}B = 12.71 \pm 0.30$‰). This suggests that modern vital effects are probably not applicable to Eocene planktonic foraminifera. Given that modern vital effects appear to be unsuitable for the Eocene, as explained above, we argue by extension that the reduced scatter in the Eocene, and notably the agreement in $\delta^{11}B_c$ among the mixed-layer species targeted in the Eocene (in stark contrast with modern values; Extended Data Fig. 7c), suggests a more minor role for vital effects in determining $\delta^{11}B$, and therefore that the targeted Eocene foraminifera must more closely reflect the $\delta^{11}B$ of ambient borate ion. Furthermore, different size fractions of the Eocene foraminifera belonging to the genus *Acarinina* from this study (Supplementary Table 1), and for Early Eocene/Palaeocene *Morozovella velascoensis* and *A. soldadoensis* in ref. 68 have $\delta^{11}B_c$ compositions within analytical uncertainty. This contrasts with the modern symbiont-bearing planktonic foraminifera *T. sacculifer* and *Globigerinoides ruber* that exhibit large changes in $\delta^{11}B$ with size (1‰ to 2.3‰; refs 64, 69), again supporting our assertion that vital effects are reduced in the Eocene foraminifera.

Although we do not yet have a full mechanistic understanding of the causes of $\delta^{11}B$ vital effects in planktonic foraminifera, a reduction in vital effects for the Eocene foraminifera would most probably result from a reduction in the relative magnitude of one or more of the following physiological processes: calcification, photosynthesis or respiration. For instance, for smaller modern foraminifera the magnitude of the influence of symbiont photosynthesis on the pH of the diffusive boundary layer is reduced, with smaller-size-fraction foraminifera recording $\delta^{11}B_c$ closer to that of seawater borate[64,69]. Here, the size of our shallower Eocene foraminifera is mostly within 212–250 μm, which is reduced compared to the modern foraminifera analysed here (300–355 μm) and elsewhere (≥425 μm; ref. 70). Also, vital-effect offsets are pH-dependent (see Extended Data Fig. 9) for symbiont-bearing foraminifera[62–64,71]. Eocene symbiotic planktonic foraminiferal composition is 15.5–14.5‰, and for this range the *T. sacculifer* (300–355 μm; ref. 22) calibration has a minimal effect on pH estimates, as it requires only a 0.1–0.3‰ correction on $\delta^{11}B_c$ for deriving $\delta^{11}B_{borate}$.

Importantly, while the modern symbiont-bearing foraminifera calibrated to date have $\delta^{11}B$ composition offset from the seawater borate curve, only $\delta^{11}B$–pH calibrations for three dinoflagellate hosting species have so far been published[9,62–64]. Our new depth profile for the modern Indian Ocean shown in Extended Data Fig. 7c shows that some extant species (*Globigerinella siphonifera* and *Globorotalia menardii*) have minimal $\delta^{11}B_c$ offsets from seawater borate $\delta^{11}B$.

Additional support for minimal vital effects in Eocene foraminifera is also given by the relative agreement with independently calculated Eocene CO_2[5,11,19] when the borate $\delta^{11}B$–pH calibration is used (Extended Data Fig. 4). Nonetheless, we also apply the modern *T. sacculifer* calibration, because this provides an upper limit on the likely extent of $\delta^{11}B$ vital effects in Eocene foraminifera. Notably, our calculated pH and CO_2 estimates for both approaches are largely within uncertainty (Extended Data Fig. 4), highlighting that such assumptions have little influence on the conclusions drawn (Fig. 4 and Extended Data Fig. 5).

Constraints on $\delta^{11}B_{sw}$. Geochemical box models[58] and existing records of benthic foraminiferal $\delta^{11}B_c$, coupled with assumptions regarding deep-water pH evolution[72], indicate that $\delta^{11}B_{sw}$ is unlikely to have changed from its modern value by more than 2.5‰ since the EECO (with a likely maximum rate of change of 0.1‰ per million years)[58]. Here, we use new, empirically constrained, upper and lower bounds on

© 2016 Macmillan Publishers Limited. All rights reserved

Eocene $\delta^{11}B_{sw}$ from several independent approaches. The underlying methodology for quantifying a maximum and minimum $\delta^{11}B_{sw}$ is based on the observation that the $\delta^{11}B$ versus pH is nonlinear (Extended Data Fig. 2), and its curvature is affected by $\delta^{11}B_{sw}$; when $\delta^{11}B_{sw}$ is higher, for any given $\delta^{11}B_c$ gradient the reconstructed pH change is larger than when $\delta^{11}B_{sw}$ is lower. Furthermore, the curvature in the $\delta^{11}B$–pH relationship results in there being a pH below which the $\delta^{11}B$ system becomes unresponsive. The lowest foraminiferal $\delta^{11}B_c$ recorded in this study therefore provides the first constraint on the $\delta^{11}B_{sw}$, which then places an additional upper bound on $\delta^{11}B_{sw}$. The lowest $\delta^{11}B_c$ values we recovered are for *Subbotina crociapertura* of 11.54 ± 0.29‰ at 45.6 Myr ago, and *S. eocaena* of 11.97 ± 0.22‰ at 40.3 Myr ago. If $\delta^{11}B_{sw} \geq 39.5$‰, these samples do not give a valid mathematical solution for the $\delta^{11}B$ versus pH relationship, so Eocene $\delta^{11}B_{sw}$ must be lower than the modern (39.6‰; ref. 59).

The lower bound for $\delta^{11}B_{sw}$ is constrained using the modern pH gradient (corrected for anthropogenic dissolved inorganic carbon) within the top ~300 m of the water column in the tropics and assuming that this represents the probable minimum pH gradient in Eocene. This assumption is justified by (1) a positive linear relationship between $\delta^{13}C$ and pH in modern[22] and Eocene seawater (Extended Data Fig. 10 based on cGENIE output[73]), and (2) larger $\delta^{13}C$ gradients between surface and deep-dwelling Eocene planktonic foraminifera than in the modern because of enhanced metabolic rates due to warmer ocean temperatures[25]. As a result, Eocene pH gradients were most probably larger than the modern equivalent and therefore our approach is conservative. Modern tropics, in non-upwelling regions, have an average pH gradient of the first 300 m equal to 0.17 (ref. 74) with an anthropogenic CO_2 correction resulting in an uncertainty of 0.003 pH units (GLODAP v1.1; ref. 75). The 53.2-Myr-ago timeslice is used for this treatment because it has full upper-water-column depth coverage and the smallest $\delta^{11}B_c$ gradient (hence it will provide the smallest pH difference and thus the minimum $\delta^{11}B_{sw}$ value). A Monte Carlo propagation of uncertainties in pH, temperature, and composition of seawater results in the pH difference (assuming $\delta^{11}B_c = \delta^{11}B_{borate}$) between the shallowest and the deepest planktonic dwellers (ΔpH) at 53.2 Myr ago falling below 0.17 (the modern equivalent) when $\delta^{11}B_{sw} < 38.2$‰ (with 86% of simulations resulting in $\geq$0.17 ΔpH). Therefore the minimum $\delta^{11}B_{sw}$ is calculated at 38.2‰ for the whole Eocene. This exercise was repeated using the alternative methodology for dealing with $\delta^{11}B$ vital effects (assuming $\delta^{11}B_c = \delta^{11}B_{borate}$ for the deeper-dwelling non-symbiotic foraminifera, which have $\delta^{11}B_c$ within error of co-occurring *Cibicicoides*, and assuming that the shallowest foraminifera has the $\delta^{11}B$–pH calibration of modern *T. sacculifer*). The minimum constraint from 86% of potential Monte Carlo solutions is then 38.6‰.

To further constrain the maximum Eocene $\delta^{11}B_{sw}$ we used restrictions offered by estimates of Apparent Oxygen Utilization (AOU), adapted from previous approaches[7]. AOU describes the amount of dissolved oxygen consumed by the remineralization of sinking organic carbon between a selected depth and saturated surface waters. The surface seawater oxygen concentration—that is, $[O_{2\text{-}sw}]_{sat}$—depends on SST as a result of Henry's law. Here, $\delta^{18}O$-derived SSTs are ~30–33 °C (Supplementary Table 1), so corresponding $[O_{2\text{-}sw}]_{sat}$ values are ~191–182 μmol kg^{-1} (ref. 76), with potential mean oversaturation of ~13 μmol kg^{-1} (low-latitude maximum: 0.2–0.4 ml per litre[77,78]). To calculate the upper estimate of $\delta^{11}B_{sw}$, the minimum SST of the Eocene time slices is selected, and thus the maximum $[O_{2\text{-}sw}]_{sat}$ is calculated to be 204 μmol kg^{-1}. Additionally, we assume Eocene planktonic foraminifera have an oxygen minimum tolerance threshold of 44.7 μmol kg^{-1}, analogous to the low-oxygen-accustomed *Hastigerinella digitata*[79] in the modern Monterey Bay (California). Therefore, we assume Eocene deeper-dwelling planktonic foraminifera were bathed in waters with oxygen concentrations equal to or higher than the minimum oxygen level of *H. digitata* (AOU = 204 μmol kg^{-1} minus 44.7 μmol kg^{-1} = 159.3 μmol kg^{-1}). To calculate AOU for each time slice studied, we modified $\delta^{11}B_{sw}$ (similarly following the Monte Carlo approach explained above) from the minimum constraint of 38.2‰ to modern seawater (39.6‰; ref. 59). We then first calculate the difference between the calculated (initially assuming that $\delta^{11}B_c = \delta^{11}B_{borate}$) dissolved inorganic carbon at the depth occupied by our deepest-dwelling foraminifera (average), and compare this to the dissolved inorganic carbon of the shallower-dwelling (within ~50 m of the surface on average) planktonic foraminifera. The difference in dissolved inorganic carbon is then transformed into AOU using Redfield ratios (106:138, CO_2:O_2). The AOU is $\leq$159.3 μmol kg^{-1} when $\delta^{11}B_{sw}$ is at maximum 38.5‰ for the 45.6-Myr-ago time slice and 38.7‰ for the 40.3-Myr-ago time slice, resulting in 86% of potential solutions with our Monte Carlo approach satisfying the AOU constraint. The same approach was used applying the *T. sacculifer* calibration for the shallowest planktonic foraminifera, resulting in a maximum $\delta^{11}B_{sw}$ of 38.8‰ and 38.9‰ for the 45.6-Myr-ago and 40.3-Myr-ago time slice respectively. To be conservative, we use the maximum of the two estimates for each calibration method. In summary, although we are calculating the lower $\delta^{11}B_{sw}$ bound at 53.2 Myr ago and the upper at 40.3 and 45.6 Myr ago, we apply this range across the whole of the study interval (53.2–36.9 Myr ago).

Second carbonate parameter. After calculating seawater pH using $\delta^{11}B_{sw}$ and $\delta^{11}B_c$, an additional carbonate parameter is required to calculate CO_2 concentrations at any given seawater salinity and temperature. Here the second parameter we use is the saturation state of calcium carbonate (surface $\Omega_{calc} = [CO_3^{2-}]/[CO_3^{2-}]_{saturation}$). Typically, for the pre-industrial tropical surface ocean Ω_{calc} is ~6, while for Eocene Tanzania waters it is ~7 (ref. 27; Extended Data Fig. 3). In support of the narrow range of potential Ω_{calc}, a variety of modelling studies of the early Cenozoic show that surface water Ω_{calc} remains within $\pm$1, essentially constant in the early Cenozoic, independent of model boundary conditions[23,80–82]. Therefore, here we assume surface Ω_{calc} was equal to 6.5 ± 1 throughout the Eocene.

Monte Carlo pH–CO_2 estimates from planktonic foraminiferal $\delta^{11}B_c$. Atmospheric CO_2 was calculated using a Monte Carlo approach to solving the relevant carbonate system equations with 10,000 iterations, deriving mean, upper and lower bounds of 95% of the simulations. We use the following constraints for our calculations: average seawater composition for $[Ca]_{sw} = 17$ mmol kg^{-1} (refs 83–87), $[Mg]_{sw} = 38 \pm 3$ mmol kg^{-1} (ref. 84) (see also Supplementary Table 2), salinity 35 ± 2, and temperature $\pm$2 °C (with temperature derived from $\delta^{18}O$ of the shallowest-dwelling species). All simulations are iterated assuming a Gaussian distribution of these parameters within the stated 2 s.d. A similar distribution is also applied to the $\delta^{11}B_c$ of the warmest and thus shallowest-dwelling species for each timeslice (Supplementary Table 2). We did not have sufficient shallow dwellers in the 36.9-Myr-ago time slice for boron isotope work. Thus, the intermediate dweller *Turborotalia cerroazulensis* was used. *T. cerroazulensis* was also analysed in the 40.3-Myr-ago time slice, and its $\delta^{11}B_c$ and $\delta^{18}O$ offset (and uncertainty) from *Morozovelloides lehneri* (the shallowest dweller in the same timeslice) was applied to estimate the shallowest temperature at 36.9 Myr ago (Supplementary Table 2). The temperature offsets between species beyond the $\pm$2 °C uncertainty applied to our calculations are considered unlikely. Note that a Gaussian distribution is not applicable to $\delta^{11}B_{sw}$ because there is equal likelihood that it falls between the minimum and maximum constraints; we therefore applied a uniform 'flat' probability $\delta^{11}B_{sw}$ for the Monte Carlo simulations. The relative magnitude of each component of our error propagation is shown in Extended Data Fig. 4.

Correction to pH–CO_2 estimates from *T. ampliapertura* $\delta^{11}B_c$. We compared $\delta^{18}O$ measurements of the surface dwellers *Planoglobanomalina* and *Pseudohastigerina* in the Eocene and early Oligocene[35,88] with that of *T. ampliapertura* used in ref. 12, and confirmed that *T. ampliapertura* gives a slightly cooler temperature than *Pseudohastigerina*. On average, when contemporaneous shallow-dwelling planktonic foraminifera are available, *T. ampliapertura* are offset by about +0.3‰ in $\delta^{18}O$ and therefore occupy a shallower water depth than *T. cerroazulensis* (whose $\delta^{18}O$ value is consistently ~0.5‰ higher than contemporaneous Eocene shallow-dwelling planktonic foraminifera). Further, considering the $\delta^{11}B$ offset between *T. cerroazulensis* and the shallow-dwelling *M. lehneri* in the 40.3-Myr-ago time slice, we apply a +1‰ correction to *T. ampliapertura* $\delta^{11}B$ values of ref. 12 to estimate surface seawater values.

Climate sensitivity calculations. Our new CO_2 reconstructions allow us to examine their role in driving long-term Eocene cooling. Climate forcing compared to the early Eocene timeslice at 53.2 Myr ago is calculated from equation (1)[28] for our two CO_2 scenarios. We then assume that the apparent latitudinal SST sensitivity for the Eocene was similar to that of the last ~520 kyr (ref. 13). Estimates of SST change per W m^{-2} forcing for each degree of latitude were generated using the second-order polynomial fit through the data of ref. 13 (dotted line in Fig. 4b, Extended Data Fig. 5b). As a high-latitude SST sensitivity we used the average of estimates for 60° N and 60° S (1.63 K per W m^{-2}); as a low-latitude estimate we used the average of values between 30° S and 30° N (0.36 K per W m^{-2}). We use the ratio of high- or low-latitude SST sensitivity and forcing for each timeslice to calculate the SST change expected as a result of our reconstructed CO_2 forcing. The product gives the estimated SST change of Fig. 4c and Extended Data Fig. 5c for high (blue symbols) and low (red symbols) latitudes driven by CO_2. The uncertainties (Fig. 4c, Extended Data Fig. 5c) are the product of 1,000 realizations of each CO_2 estimate within its uncertainty envelope (95% confidence intervals).

To test the importance of CO_2 to early and late Eocene warmth, we calculated the sensitivity of the Earth system to radiative forcing, the ECS. For ΔT we use global mean surface temperature for the EECO and late Eocene, where the EECO ΔT is thought to be ~14 $\pm$ 3 °C warmer than the pre-industrial global mean surface temperature, and ~5 °C warmer than the late Eocene global mean surface temperature (35 Myr ago)[2,29]. Of this ΔT, 4–6 °C is attributed to changes in boundary conditions (palaeotopography and continental configuration of the Eocene) and the action of the slow climate feedbacks (mainly the lack of a continental ice sheet on Antarctica and vegetation change[2–4]), for both the early and late Eocene. The uncertainties in Fig. 4d to 4f (and Extended Data Fig. 5d–f) are the products of 1,000 realizations of the 53.2-Myr-ago and 36.9-Myr-ago CO_2 values and the ΔT

© 2016 Macmillan Publishers Limited. All rights reserved

estimate based on randomly sampling each variable within its 95% confidence interval uncertainty envelope. The uncertainty for the CO_2 and ΔT estimates follows a Gaussian distribution, and for the slow-feedback correction on ΔT follows a uniform 'flat' probability.

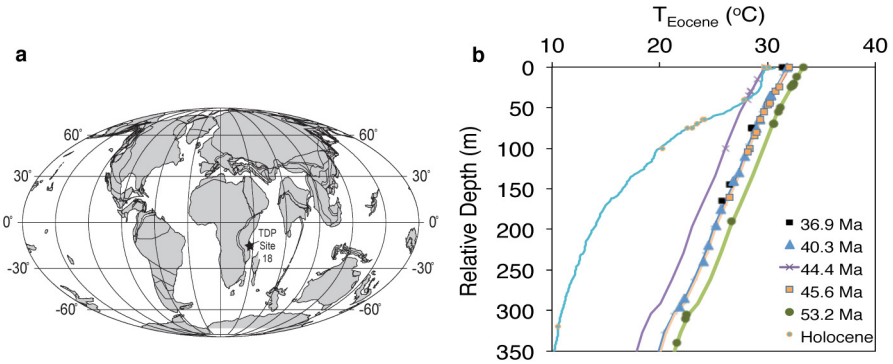
Data. The underlying data of this manuscript may be found in the Supplementary Tables 1 and 2 (refs 91–94).

31. Nicholas, C. J. *et al.* Stratigraphy and sedimentology of the Upper Cretaceous to Paleogene Kilwa Group, southern coastal Tanzania. *J. Afr. Earth Sci.* **45,** 431–466 (2006).
32. Pearson, P. N. *et al.* Further Paleogene and Cretaceous sediment cores from the Kilwa area of coastal Tanzania: Tanzania Drilling Project Sites 6-10. *J. Afr. Earth Sci.* **45,** 279–317 (2006).
33. Kent, P. E., Hunt, J. A. & Johnstone, D. W. The Geology and Geophysics of Coastal Tanzania. Geophys. Paper No. 6, 1–101 (Her Majesty's Stationery Office, Inst. Geol. Sci, 1971).
34. Bown, P. R. *et al.* A Paleogene calcareous microfossil Konservat-Lagerstätte from the Kilwa Group of coastal Tanzania. *Geol. Soc. Am. Bull.* **120,** 3–12 (2008).
35. Pearson, P. N. *et al.* Stable warm tropical climate through the Eocene Epoch. *Geology* **35,** 211–214 (2007).
36. van Dongen, B. E. *et al.* Well preserved Palaeogene and Cretaceous biomarkers from the Kilwa area, Tanzania. *Org. Geochem.* **37,** 539–557 (2006).
37. Takahashi, T. *et al.* Climatological mean and decadal change in surface ocean pCO_2, and net sea-air CO_2 flux over the global oceans. *Deep Sea Res. II* **56,** 554–577 (2009).
38. Barker, S., Greaves, M. & Elderfield, H. A study of cleaning procedures used for foraminiferal Mg/Ca paleothermometry. *Geochem. Geophys. Geosyst.* **4,** 8407 (2003).
39. Yu, J., Elderfield, H., Greaves, M. & Day, J. Preferential dissolution of benthic foraminiferal calcite during laboratory reductive cleaning. *Geochem. Geophys. Geosys.* **8,** Q06016 (2007).
40. Rae, J. W. B., Foster, G. L., Schmidt, D. N. & Elliott, T. Boron isotopes and B/Ca in benthic foraminifera: proxies for the deep ocean carbonate system. *Earth Planet. Sci. Lett.* **302,** 403–413 (2011).
41. Edgar, K. M., Anagnostou, E., Pearson, P. N. & Foster, G. L. Assessing the impact of diagenesis on $\delta^{11}B$, $\delta^{13}C$, $\delta^{18}O$, Sr/Ca and B/Ca values in fossil planktic foraminiferal calcite. *Geochim. Cosmochim. Acta* **166,** 189–209 (2015).
42. Kim, S.-T. & O'Neil, J. R. Equilibrium and non-equilibrium oxygen isotope effects in synthetic carbonates. *Geochim. Cosmochim. Acta* **61,** 3461–3475 (1997).
43. Cramer, B. S., Miller, K. G., Barrett, P. J. & Wright, J. D. Late Cretaceous–Neogene trends in deep ocean temperature and continental ice volume: reconciling records of benthic foraminiferal geochemistry ($\delta^{18}O$ and Mg/Ca) with sea level history. *J. Geophys. Res.* **116,** C12023 (2011).
44. Zachos, J. C., Stott, L. D. & Lohmann, K. C. Evolution of early Cenozoic marine temperatures. *Paleoceanography* **9,** 353–387 (1994).
45. Pearson, P. N. in *Reconstructing Earth's Deep Time Climate—The State of the Art in 2012* Vol. 18 (eds Ivany, L. C. & Huber, B. T.) 1–38 (The Paleontological Society, 2012).
46. LeGrande, A. N. & Schmidt, G. A. Global gridded data set of the oxygen isotopic composition in seawater. *Geophys. Res. Lett.* **33,** L12604 (2006).
47. Sexton, P. F., Wilson, P. A. & Pearson, P. N. Palaeoecology of late middle Eocene planktic foraminifera and evolutionary implications. *Mar. Micropaleontol.* **60,** 1–16 (2006).
48. Pearson, P. N., Shackleton, N. J. & Hall, M. A. Stable isotope paleoecology of middle Eocene planktonic foraminifera and multi-species isotope stratigraphy, DSDP Site 523, South Atlantic. *J. Foraminiferal Res.* **23,** 123–140 (1993).
49. Pearson, P. N. *et al.* Warm tropical sea surface temperatures in the Late Cretaceous and Eocene epochs. *Nature* **413,** 481–487 (2001).
50. Edgar, K. M. *et al.* Symbiont 'bleaching' in planktic foraminifera during the Middle Eocene Climatic Optimum. *Geology* **41,** 15–18 (2013).
51. Birch, H. *et al.* Planktonic foraminifera stable isotopes and water column structure: disentangling ecological signals. *Mar. Micropaleontol.* **101,** 127–145 (2013).
52. Fairbanks, R. G. & Wiere, P. H. & Bé, A. W. H. Vertical distribution and isotopic composition of living planktonic foraminifera in the western North Atlantic. *Science* **207,** 61–63 (1980).
53. Spero, H. & Williams, D. F. Extracting environmental information from planktonic foraminiferal $\delta^{13}C$ data. *Nature* **335,** 717–719 (1988).
54. Berger, W. H., Killingley, J. S. & Vincent, E. Stable isotopes in deep-sea carbonates: Box Core ERDC-92, West Equatorial Pacific. *Oceanol. Acta* **1,** 203–216 (1978).
55. Seki, O. *et al.* Alkenone and boron-based Pliocene pCO_2 records. *Earth Planet. Sci. Lett.* **292,** 201–211 (2010).
56. Hönisch, B. *et al.* Atmospheric carbon dioxide concentration across the Mid-Pleistocene transition. *Science* **324,** 1551–1554 (2009).
57. Sanyal, A. & Bijma, J. A comparative study of the northwest Africa and eastern equatorial Pacific upwelling zones as sources of CO_2 during glacial periods based on boron isotope paleo-pH estimation. *Paleoceanography* **14,** 753–759 (1999).
58. Lemarchand, D., Gaillardet, J., Lewin, E. & Allegre, C. J. The influence of rivers on marine boron isotopes and implications for reconstructing past ocean pH. *Nature* **408,** 951–954 (2000).
59. Foster, G. L., Pogge von Strandmann, P. A. E. & Rae, J. W. B. Boron and magnesium isotopic composition of seawater. *Geochem. Geophys. Geosyst.* **11,** Q08015 (2010).
60. Rink, S., Kühl, M., Bijma, J. & Spero, H. J. Microsensor studies of photosynthesis and respiration in the symbiotic foraminifer *Orbulina universa*. *Mar. Biol.* **131,** 583–595 (1998).
61. Zeebe, R. E., Wolf-Gladrow, D. A., Bijma, J. & Hönisch, B. Vital effects in foraminifera do not compromise the use of $\delta^{11}B$ as a paleo-pH indicator: evidence from modelling. *Paleoceanography* **18,** 1043 (2003).
62. Sanyal, A. *et al.* Oceanic pH control on the boron isotopic composition of foraminifera: evidence from culture experiments. *Paleoceanography* **11,** 513–517 (1996).
63. Sanyal, A., Bijma, J., Spero, H. & Lea, D. W. Empirical relationship between pH and the boron isotopic composition of *Globigerinoides sacculifer*: implications for the boron isotope paleo-pH proxy. *Paleoceanography* **16,** 515–519 (2001).
64. Henehan, M. J. *et al.* Calibration of the boron isotope proxy in the planktonic foraminifera *Globigerinoides ruber* for use in palaeo-CO_2 reconstruction. *Earth Planet. Sci. Lett.* **364,** 111–122 (2013).
65. Yu, J., Thornalley, D. J. R., Rae, J. W. B. & McCave, N. I. Calibration and application of B/Ca, Cd/Ca, and $\delta^{11}B$ in *Neogloboquadrina pachyderma* (sinistral) to constrain CO_2 uptake in the subpolar North Atlantic during the last deglaciation. *Paleoceanography* **28,** 237–252 (2013).
66. Kroon, D. & the Shipboard Scientific Party. Tropical Temperature History during Paleogene Global Warming (GLOW) Events. Netherlands Institute for Sea Research (NIOZ) Site Survey Cruise Report (RV Pelagia cruise number 64PE303) 1–151 (NIOZ, 2010).
67. McCorkle, D., Corliss, B. & Farnham, C. Vertical distributions and stable isotopic compositions of live (stained) benthic foraminifera from the North Carolina and California continental margins. *Deep Sea Res. I* **44,** 983–1024 (1997).
68. Penman, D. E. *et al.* Rapid and sustained surface ocean acidification during the Paleocene-Eocene Thermal Maximum. *Paleoceanography* **29,** 2014PA002621 (2014).
69. Hönisch, B. & Hemming, N. G. Ground-truthing the boron isotope-paleo-pH proxy in planktonic foraminifera shells: partial dissolution and shell size effects. *Paleoceanography* **19,** PA4010 (2004).
70. Martínez-Boti, M. A. *et al.* Boron isotope evidence for oceanic carbon dioxide leakage during the last deglaciation. *Nature* **518,** 219–222 (2015).
71. Hönisch, B. *et al.* The influence of symbiont photosynthesis on the boron isotopic composition of foraminifera shells. *Mar. Micropaleontol.* **49,** 87–96 (2003).
72. Raitzsch, M. & Hönisch, B. Cenozoic boron isotope variations in benthic foraminifers. *Geology* **41,** 591–594 (2013).
73. John, E. H., Wilson, D. J., Ridgwell, A. & Pearson, P. N. Temperature-dependent remineralization and carbon cycling in the warm Eocene oceans. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **413,** 158–166 (2014).
74. Goyet, C., Healy, R. J. & Ryan, J. P. Global distribution of total inorganic carbon and total alkalinity below the deepest winter mixed layer depths. Report ORNL/CDIAC-127, NDP-076 (Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, US Department of Energy, 2000).
75. Key, R. M. *et al.* A global ocean carbon climatology: results from the Global Data Analysis Project (GLODAP). *Glob. Biogeochem. Cycles* **18,** GB4031 (2004).
76. Battino, R., Rettich, T. R. & Tominaga, T. The solubility of oxygen and ozone in liquids. *J. Phys. Chem. Ref. Data* **12,** 163–178 (1983).
77. Jenkins, W. J. & Goldman, J. C. Seasonal oxygen cycling and primary production in the Sargasso Sea. *J. Mar. Res.* **43,** 465–491 (1985).
78. Sarmiento, J. L. & Gruber, N. *Ocean Biogeochemical Dynamics* 1–528 (Princeton Univ. Press, 2006).
79. Hull, P. M., Osborn, K. J., Norris, R. D. & Robinson, B. H. Seasonality and depth distribution of a mesopelagic foraminifer, *Hastigerinella digitata*, in Monterey Bay, California. *Limnol. Oceanogr.* **56,** 562–576 (2011).
80. Pälike, H. *et al.* A Cenozoic record of the equatorial Pacific carbonate compensation depth. *Nature* **488,** 609–614 (2012).
81. Ridgwell, A. & Zeebe, R. E. The role of the global carbonate cycle in the regulation and evolution of the Earth system. *Earth Planet. Sci. Lett.* **234,** 299–315 (2005).
82. Tyrrell, T. & Zeebe, R. E. History of carbonate ion concentration over the last 100 million years. *Geochim. Cosmochim. Acta* **68,** 3521–3530 (2004).
83. Coggon, R. M. *et al.* Reconstructing past seawater Mg/Ca and Sr/Ca from mid-ocean ridge flank calcium carbonate veins. *Science* **327,** 1114–1117 (2010).
84. Horita, J., Zimmermann, H. & Holland, H. D. Chemical evolution of seawater during the Phanerozoic: implications for the record of marine evaporites. *Geochim. Cosmochim. Acta* **66,** 3733–3756 (2002).
85. Dickson, A. J. Fossil echinoderms as monitor of the Mg/Ca ratio of Phanerozoic oceans. *Science* **298,** 1222–1224 (2002).
86. Lowenstein, T. K. *et al.* Oscillations in the Phanerozoic seawater chemistry: evidence from fluid inclusions. *Science* **294,** 1086–1088 (2001).
87. Lear, C. H., Elderfield, H. & Wilson, P. A. Cenozoic deep-sea temperatures and global ice volumes from Mg/Ca in benthic foraminiferal calcite. *Science* **287,** 269–272 (2000).
88. Wade, B. S. & Pearson, P. N. Planktonic foraminiferal turnover, diversity fluctuations and geochemical signals across the Eocene/Oligocene boundary in Tanzania. *Mar. Micropaleontol.* **68,** 244–255 (2008).
89. Huber, M. & Caballero, R. The early Eocene equable climate problem revisited. *Clim. Past* **7,** 603–633 (2011).

© 2016 Macmillan Publishers Limited. All rights reserved

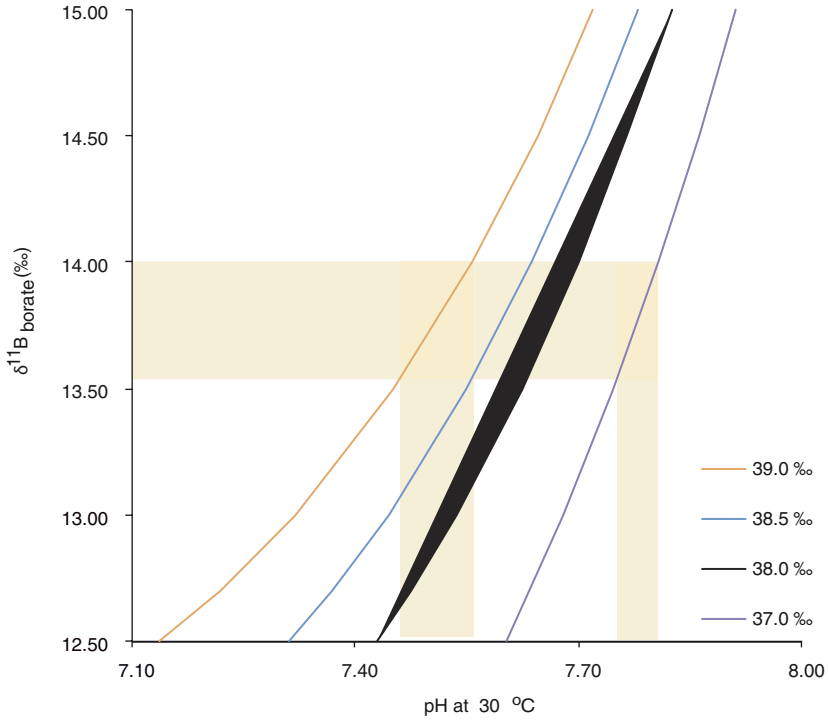
90. Tindall, J. *et al.* Modelling the oxygen isotope distribution of ancient seawater using a coupled ocean-atmosphere GCM: implications for reconstructing early Eocene climate. *Earth Planet. Sci. Lett.* **292,** 265–273 (2010).
91. Wade, B. S., Pearson, P. N., Berggren, W. A. & Pälike, H. Review and revision of Cenozoic tropical planktonic foraminiferal biostratigraphy and calibration to the geomagnetic polarity and astronomical time scale. *Earth Sci. Rev.* **104,** 111–142 (2011).
92. Ben-Yaakov, S. & Goldhaber, M. B. The influence of sea water composition on the apparent constants of the carbonate system. *Deep-Sea Res.* **20,** 87–99 (1973).
93. Millero, F. J. & Pierrot, D. A chemical equilibrium model for natural waters. *Aquat. Geochem.* **4,** 153–199 (1998).
94. Hershey, J. P., Fernandez, M., Milne, P. J. & Millero, F. J. The ionization of boric acid in NaCl, Na-Ca-Cl and Na-Mg-Cl solutions at 25°C. *Geochim. Cosmochim. Acta* **50,** 143–148 (1986).

© 2016 Macmillan Publishers Limited. All rights reserved



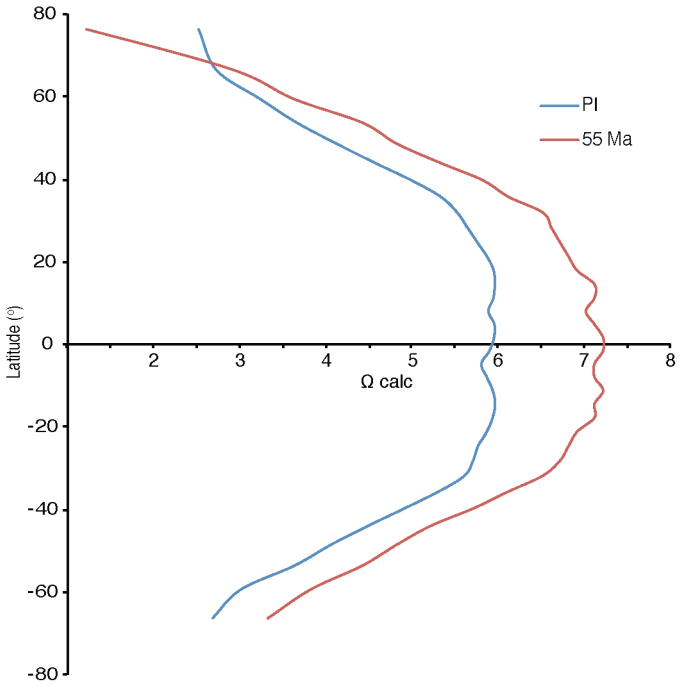
Extended Data Figure 1 | Palaeogeography and $\delta^{18}O$-derived temperature against foraminiferal calcification depth. a, Approximate palaeoposition of Tanzanian Drilling Project (TDP) sites studied here (map generated from www.odsn.de). **b**, Reconstructed temperature (T_{Eocene}) and relative depth of each foraminifera within each time slice. The pale blue line represents the output of a General Circulation Model simulation run with Eocene boundary conditions[89,90], whereas the other coloured lines show the General Circulation Model output offset to intersect with the warmest temperature at depth zero for each time slice. Note that depth assignments are approximate (see Methods). Also shown are the reconstructed temperatures from core-top (Holocene) planktonic foraminiferal $\delta^{18}O$ (Supplementary Table 1), where *G. ruber* is assigned a depth of zero and the rest of the planktonic foraminifera are offset to reproduce the measured temperature profile of the GLOW15 cruise (light blue/green line)[51,66]. For Holocene temperature reconstructions we used the modern site latitude of 9° S, and $\delta^{18}O_{sw}$ (using the Standard Mean Ocean Water (SMOW) standard) of 0‰.

© 2016 Macmillan Publishers Limited. All rights reserved



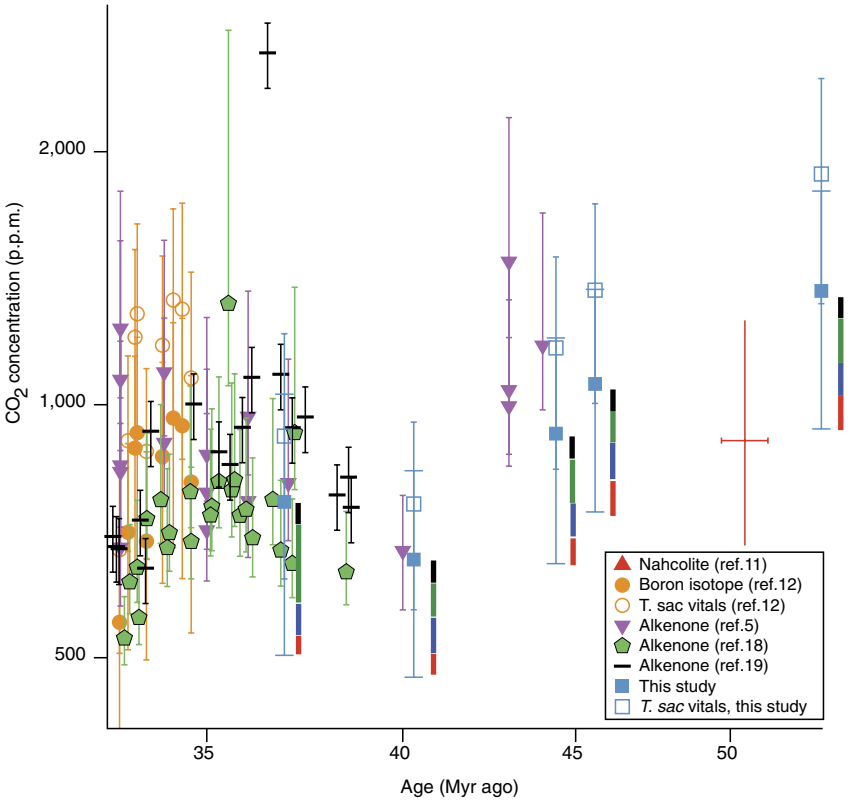
Extended Data Figure 2 | $\delta^{11}B$ versus pH as a function of $\delta^{11}B_{sw}$. Increasing $\delta^{11}B_{sw}$, as indicated for each line in units of per mil, results in lower pH for the same $\delta^{11}B$. However, for the same $\delta^{11}B$ range, the reconstructed pH range is larger for higher $\delta^{11}B_{sw}$ (see brown-shaded regions).

© 2016 Macmillan Publishers Limited. All rights reserved



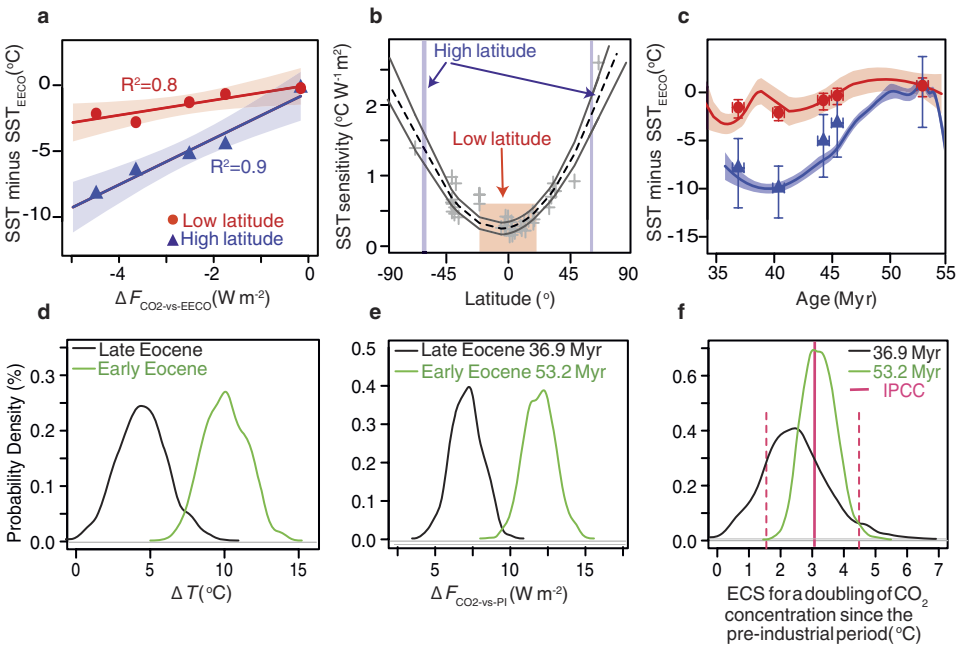
Extended Data Figure 3 | cGENIE estimates of calcite saturation in surface waters. Comparison of calcite saturation Ω_{calc} for pre-industrial times (PI; blue) and 55 Myr ago (red) at different latitudes.

© 2016 Macmillan Publishers Limited. All rights reserved



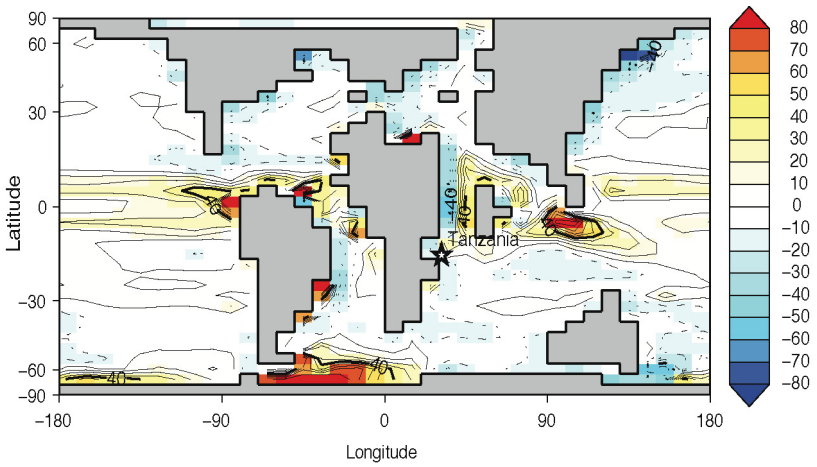
Extended Data Figure 4 | Compilation of several CO_2 records for the Eocene in comparison to this study. Data are from refs 5, 11, 12, 18 and 19. The contribution of different parameters to the uncertainty on our CO_2 reconstructions is colour-coded; sequentially from bottom to top, red is from $\delta^{11}B_{sw}$, blue is from the Ω_{calc} uncertainty, green is from the $\delta^{11}B_c$ error, black is the 40 p.p.m. uncertainty in the event of disequilibrium with the atmosphere. Other parameters contribute <10% uncertainty to the CO_2 calculations and are not shown. Note that the data from Pearson *et al.*[12] (orange circles) are corrected as in Fig. 3, and there are two scenarios included for the Tanzania records of ref. 12 and this study: one with $\delta^{11}B_c = \delta^{11}B_{borate}$ (closed blue and orange symbols) and the other applying *T. sacculifer* (open blue and orange symbols) corrections to the shallowest symbiotic planktonic foraminifera (Methods). The *y* axis is in log-scale. Error bars are representative of each proxy's reconstruction uncertainty (typically at 95% confidence). For the $\delta^{11}B$ reconstructions in this study, the errors are based on Monte Carlo propagation of relevant errors (Methods).

© 2016 Macmillan Publishers Limited. All rights reserved



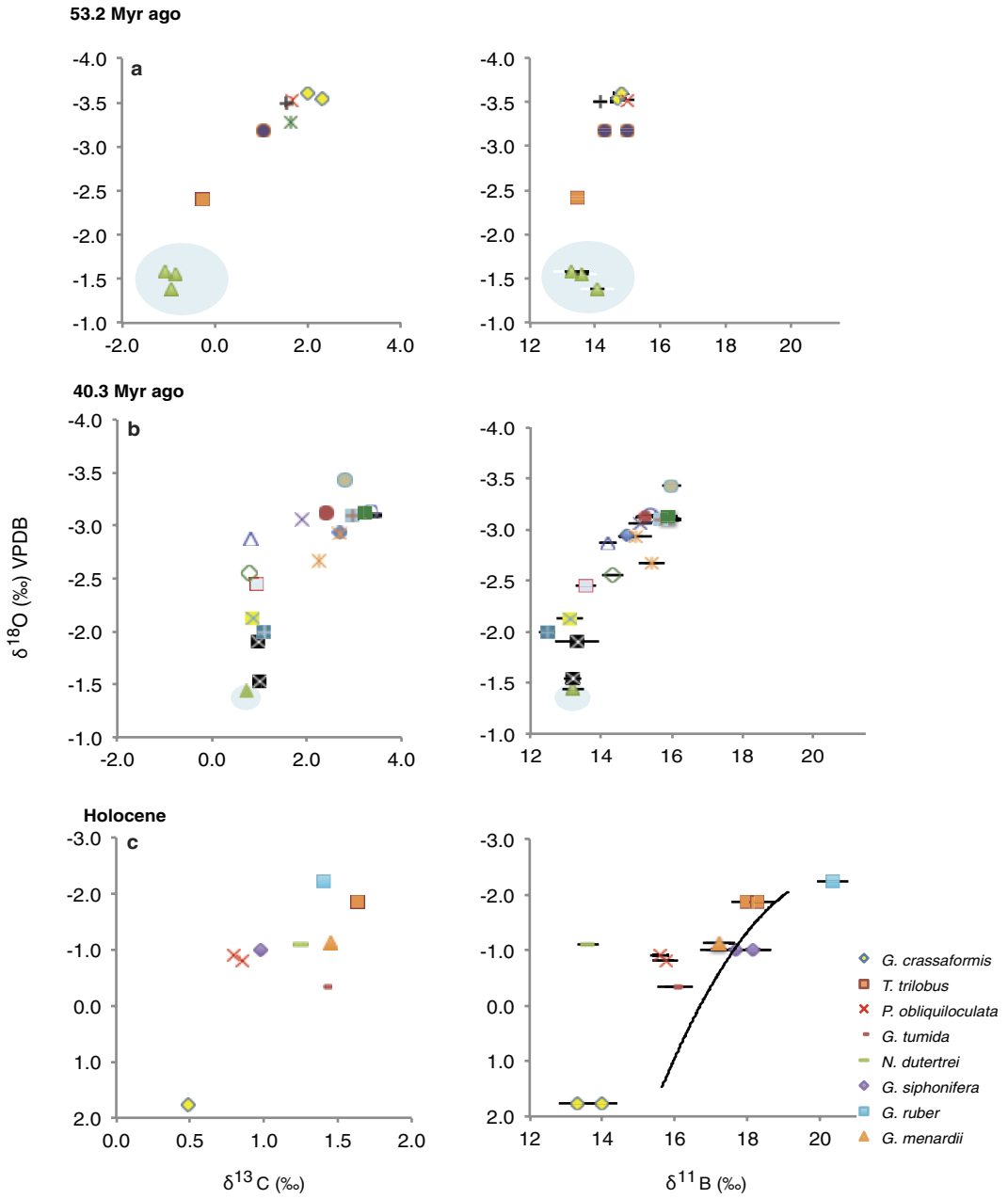
Extended Data Figure 5 | CO_2 as a driver of latitudinal cooling in the Eocene, and ECS analyses of the EECO and late Eocene time slices. The case for the *T. sacculifer* calibration applied to shallowest planktonic foraminifera. **a**, Evolving relationship between SST[14] for high (blue) and low (red) latitudes and the CO_2 forcing of each of our time slices relative to the EECO. Linear regression fits and coefficients of determination (R^2) are also shown, with the 95% confidence interval (shaded bands). **b**, Apparent latitudinal SST sensitivity for the last 520 kyr (ref. 13). The dashed line is a second-order polynomial through the SST sensitivity data (grey crosses) of ref. 13, and the grey lines show the 95% confidence interval. A red rectangle surrounds the SST sensitivity estimates averaged as a low-latitude mean, and the blue line indicates the high-latitude mean (see text and Methods). **c**, Reconstructed (lines) and estimated (symbols) SST relative to 53.2 Myr ago. Symbols show each of our time slices, calculated using the respective CO_2 reconstructions and the average low- (red) and high- (blue) latitude SST sensitivities of **b**. Bold lines and shaded uncertainty band (at 95% confidence) show the reconstructed long-term mean SST estimates using the TEX_{86} proxy at high (blue colour >55°) and low (red colour <30°) latitudes[14] relative to the SST ~53 Myr ago. Error bars represent full propagation of errors at 95% confidence[14]. **d**, Range in mean surface temperature change for early (green) and late (black) Eocene corrected for changes due to slow feedbacks[2,4,21,29]. **e**, Forcing compared to the pre-industrial period, calculated using our CO_2 reconstructions for the time slices 53.2 Myr ago (early Eocene) and 36.9 Myr ago (late Eocene) (see Methods). **f**, Probability density functions of ECS for the early (green) and late (black) Eocene compared to IPCC estimates (dashed lines show the 95% confidence interval (solid pink line)).

© 2016 Macmillan Publishers Limited. All rights reserved



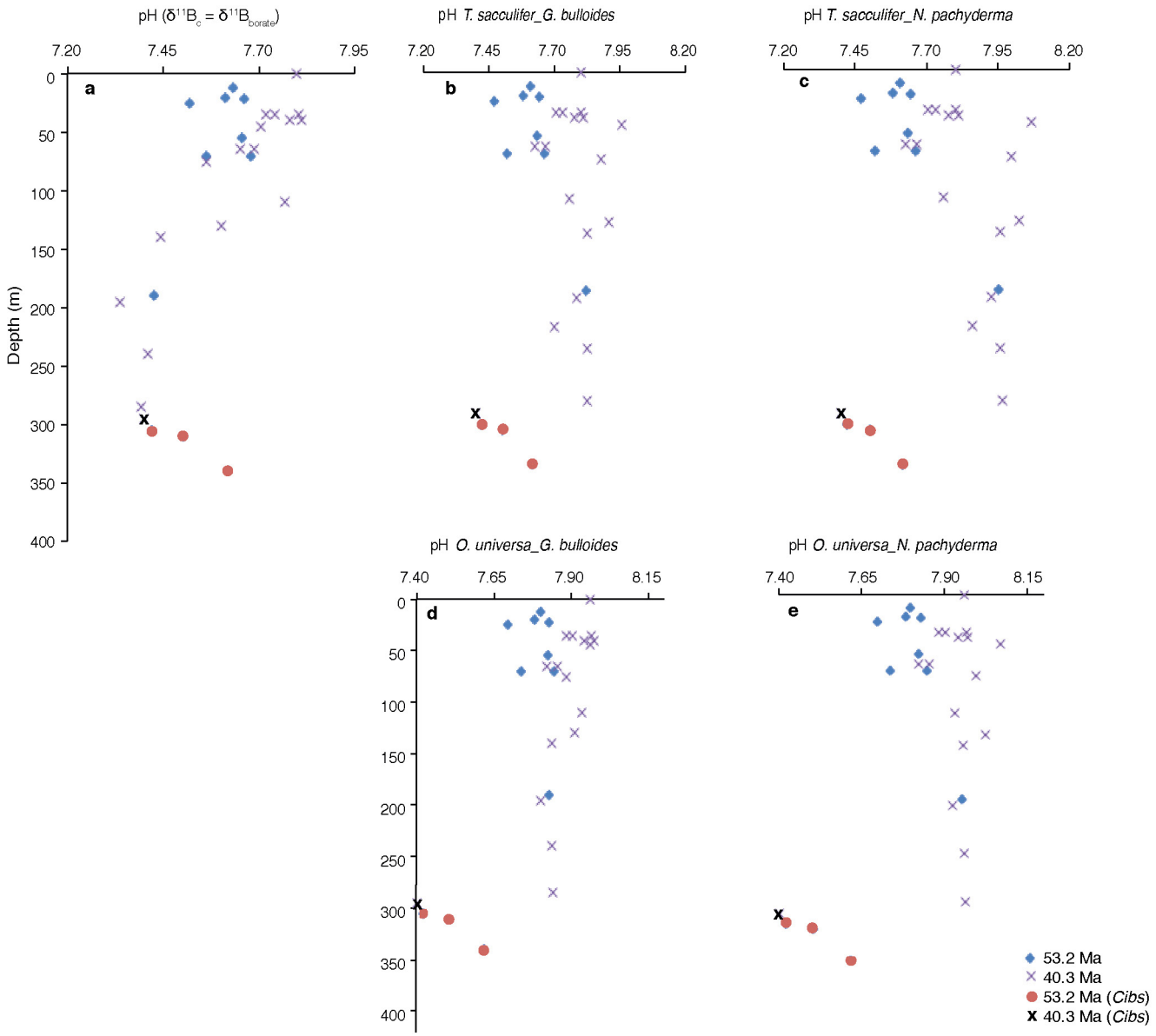
Extended Data Figure 6 | The cGENIE estimates of air–seawater CO_2 disequilibrium. The colour scale shows the difference between pCO_2 in air and pCO_2 in sea water. The model uses Eocene boundary conditions and positive values mean that seawater is a source of CO_2 (in parts per million); the star shows the palaeo-location of Tanzania.

© 2016 Macmillan Publishers Limited. All rights reserved



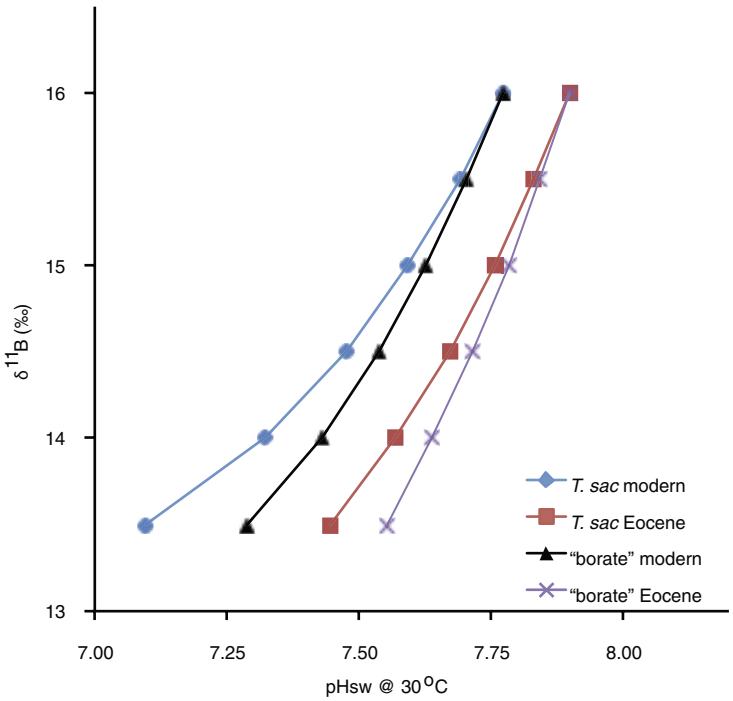
Extended Data Figure 7 | Comparison of Eocene and modern planktonic foraminiferal $\delta^{11}B$ and $\delta^{13}C$ with $\delta^{18}O$. a and **b** show analyses from the time slices 53.2 Myr ago and 40.3 Myr ago, respectively (as in Fig. 2 and Supplementary Table 1). *Cibicidoides* species (*Cibs*) are shaded in blue. **c**, Core-top (Holocene) offshore Tanzania foraminiferal measurements. Seawater $\delta^{11}B_{borate}$ and $\delta^{18}O$ was calculated from temperature, alkalinity and dissolved inorganic carbon measurements (from GLODAP cruises 18 and 23, stations 17742 and 23037, and 53.96° E to 7.04° S and 52.37° E to 6.33° S, respectively), correcting for anthropogenic carbon input. The black line in **c** represents seawater-derived $\delta^{18}O$ and $\delta^{11}B_{borate}$ data (Methods). The symbols for the time slices 53.2 and 40.3 Myr ago are as in Fig. 2. Note the change in scale for the *x* axis between the Eocene and Holocene panels. Errors in $\delta^{11}B$ represent 2 s.d. of long-term precision (Methods).

© 2016 Macmillan Publishers Limited. All rights reserved



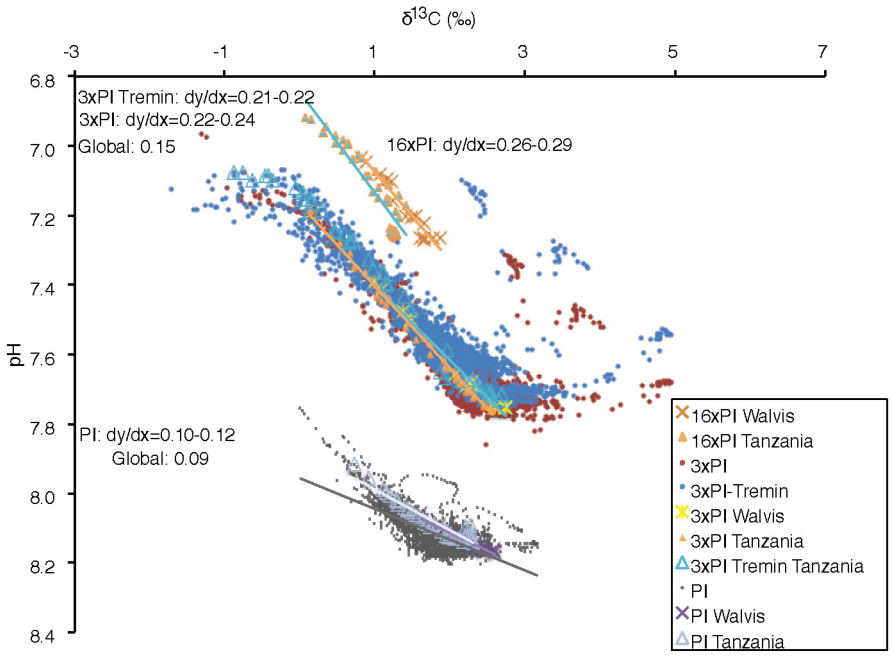
Extended Data Figure 8 | Reconstructed pH using different combinations of published symbiont-bearing and non-symbiotic foraminiferal vital-effect calibrations for the two most complete with depth time slices. a shows the case where no vital effect corrections were applied for comparison. The vital-effect corrections for deeper asymbiotic planktonic foraminifera are based on either *Globigerina bulloides*[26] (**b** and **d**) and *Neogloboquadrina pachyderma*[65] (**c** and **e**) calibrations. For shallow symbiont-bearing planktonics we used calibrations specific to modern shallow, symbiont-bearing *Trilobatus sacculifer*[22] (**b** and **c**) and *Orbulina universa*[62] (**d** and **e**) (as recalculated by ref. 64 and offset by −3‰ to account for analytical differences between negative thermal ionization mass spectrometry and MC-ICPMS instrumentation[8,9]). For comparison we also show the pH reconstructions for the case where we assume $\delta^{11}B_c = \delta^{11}B_{borate}$.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 9 | The effect of seawater composition on boron isotope calibrations in foraminifera. The example of *T. sacculifer* calibration[22] and the 'borate' calibration (assuming $\delta^{11}B_c = \delta^{11}B_{borate}$) for modern and Eocene seawater compositions; see Methods and Supplementary Table 2.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 10 | The cGENIE output of seawater pH versus $\delta^{13}C$ for the top ~300 m of the ocean. The scenarios explored are from offshore Tanzania (triangles), Walvis Ridge in the South Atlantic Ocean (crosses), and the global ocean (circles) at three different atmospheric CO_2 concentrations (modern, 3× pre-industrial and 16× pre-industrial, PI). An additional scenario showing 3× pre-industrial CO_2 but considering the temperature effect on remineralization (indicated as 'Tremin') is also shown as blue squares and triangles.

© 2016 Macmillan Publishers Limited. All rights reserved

LETTER

doi:10.1038/nature17447

Repeated large-scale retreat and advance of Totten Glacier indicated by inland bed erosion

A. R. A. Aitken[1], J. L. Roberts[2,3], T. D. van Ommen[2,3], D. A. Young[4], N. R. Golledge[5,6], J. S. Greenbaum[4], D. D. Blankenship[4] & M. J. Siegert[7]

Climate variations cause ice sheets to retreat and advance, raising or lowering sea level by metres to decametres. The basic relationship is unambiguous, but the timing, magnitude and sources of sea-level change remain unclear; in particular, the contribution of the East Antarctic Ice Sheet (EAIS) is ill defined, restricting our appreciation of potential future change. Several lines of evidence suggest possible collapse of the Totten Glacier into interior basins during past warm periods, most notably the Pliocene epoch[1–4], causing several metres of sea-level rise. However, the structure and long-term evolution of the ice sheet in this region have been understood insufficiently to constrain past ice-sheet extents. Here we show that deep ice-sheet erosion—enough to expose basement rocks—has occurred in two regions: the head of the Totten Glacier, within 150 kilometres of today's grounding line; and deep within the Sabrina Subglacial Basin, 350–550 kilometres from this grounding line. Our results, based on ICECAP aerogeophysical data, demarcate the marginal zones of two distinct quasi-stable EAIS configurations, corresponding to the 'modern-scale' ice sheet (with a marginal zone near the present ice-sheet margin) and the retreated ice sheet (with the marginal zone located far inland). The transitional region of 200–250 kilometres in width is less eroded, suggesting shorter-lived exposure to eroding conditions during repeated retreat–advance events, which are probably driven by ocean-forced instabilities. Representative ice-sheet models indicate that the global sea-level increase resulting from retreat in this sector can be up to 0.9 metres in the modern-scale configuration, and exceeds 2 metres in the retreated configuration.

Satellite-based observations indicate that the margin of Totten Glacier may be experiencing greater ice loss than anywhere else in East Antarctica[5,6]. This, coupled with the presence of low-lying subglacial basins upstream[7,8], means the Totten Glacier catchment area could be at risk of substantial ice loss under ocean-warming conditions. The vulnerability of this region to change might be driven by the entrance of warm modified circumpolar deep water to the Totten Ice Shelf cavity[9,10], for which bathymetric pathways are known[11].

Totten Glacier possesses an ice stream that extends far inland[12]. This ice stream mostly overlies the Sabrina Subglacial Basin (SSB), which is a bowl-shaped depression bounded to the east by the Terre Adelie highlands, and to the west and south by the recently discovered Highlands B and C (Fig. 1). This region has widely distributed subglacial hydrology[13] and little large-scale topographic relief[8,13,14]. The SSB is underlain by a sedimentary basin (the SSSB) with moderate but quite variable thickness (Fig. 1c). The SSSB probably dates back to at least 40 million years ago (Ma)[15], and therefore pre-dates the glaciation of the EAIS.

Here we use geophysical data from the International Collaborative Exploration of the Cryosphere through Airborne Profiling (ICECAP) program (Extended Data Fig. 1) to define the erosion of the SSB by past EAIS activity. We use two-dimensional gravity modelling along flight lines, and also include estimates of depth to magnetic sources, to understand the thickness of sedimentary rocks in the SSSB (Fig. 1c). The thickness of the SSSB is defined with a typical accuracy of ±600 metres.

The gravity-model results show that the base of the SSSB sedimentary basin is tilted towards the south–southeast and extends southeast from the Totten Glacier for more than 500 kilometres. Its base is as deep as −4,000 metres elevation in parts, but is typically shallower (Fig. 1b). The pre-erosion basin structure is not uniquely defined, as the basin top is missing, but the least-eroded areas of the SSSB are at least 3 kilometres thick and have a generally flat-based geometry. Faults define smaller-scale perturbations to this geometry. The Aurora Subglacial Basin (ASB) and Vincennes Subglacial Basin (VSB) regions have much larger thickness uncertainties, and we have not interpreted erosion patterns within these basins.

Thickness trends for the SSSB (Fig. 1c) do not parallel the tectonic structure of the region (Extended Data Fig. 2). Thickness variations are high within tectonic blocks, and thickness trends transgress major tectonic structures[15,16]. This suggests that present-day basin thickness is not dominated by tectonic structure; rather, the SSSB thickness defines distinct patterns that are explained by the erosion of the SSB by the EAIS.

Glacial erosion occurs because of basal sliding, which requires warm-based ice and a driving stress that is sufficient to cause motion[17]. The erosion rate depends primarily on the basal velocity of the ice sheet[17], and because there is the potential for high velocities at the ice-sheet margin, so there is also the potential for enhanced erosion rates at the margin. Erosion may be selective or distributed. Under selective erosion, deep troughs occur in regions with high basal velocity[17], but a lateral convergence of ice flux is also required, often focused within pre-existing valleys[18]. Thinner ice promotes selective erosion because highlands are often cold-based and do not exhibit the fast basal velocities necessary for substantial erosion; the same highlands may be warm-based under thicker ice cover, and may exhibit fast basal velocities. As a consequence, larger and thicker ice sheets are more likely to exhibit distributed erosion, whereas smaller and thinner ice sheets are more likely to exhibit selective erosion[19].

For the SSB we interpret several regions of distinctive erosion (Fig. 1d) on the basis of the morphology of the subglacial surface and the thickness of the SSSB (Extended Data Table 1). These regions define erosion from two distinct quasi-stable EAIS configurations: a 'modern-scale' configuration, with a marginal zone near the present-day ice-sheet margin; and a 'retreated' configuration, with a marginal zone located far inland.

Cumulative glacial erosion is generally low in elevated coastal regions, including Law Dome, the Knox Coast and the Cape Goodenough region, and erosion is also generally low on the highlands surrounding the SSB and ASB. Ice-sheet models suggest that these

[1]School of Earth and Environment, University of Western Australia, Perth, Western Australia 6008, Australia. [2]Australian Antarctic Division, Kingston, Tasmania 7050, Australia. [3]Antarctic Climate & Ecosystems Cooperative Research Centre, University of Tasmania, Hobart, Tasmania 7005, Australia. [4]University of Texas Institute for Geophysics, University of Texas at Austin, Austin, Texas 78758, USA. [5]Antarctic Research Centre, Victoria University of Wellington, Wellington 6140, New Zealand. [6]GNS Science, Avalon, Lower Hutt 5011, New Zealand. [7]The Grantham Institute and Department of Earth Science and Engineering, Imperial College London, London SW7 2AZ, UK.

© 2016 Macmillan Publishers Limited. All rights reserved

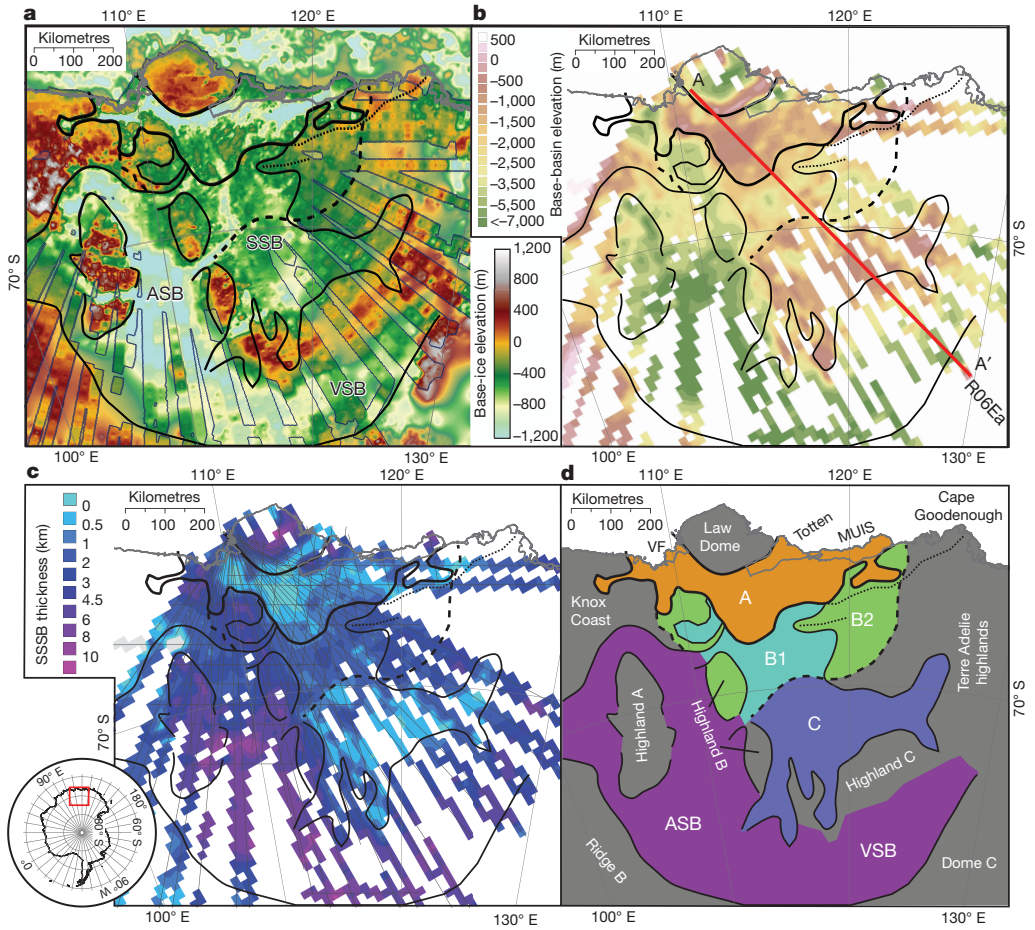


Figure 1 | Interpretation of erosion in the Sabrina Subglacial Basin region. The Sabrina Subglacial Basin (SSB) is the low-lying region bounded by Highland B, Highland C and the Terre Adelie highlands. The SSSB is the associated sedimentary basin of moderate thickness. Much thicker sedimentary basins occur in the Aurora Subglacial Basin (ASB) and Vincennes Subglacial Basin (VSB) regions. **a**, Base-ice elevations indicate the ice-sheet bed. Data are from ICECAP and Bedmap2 (ref. 7) (the latter data have a muted appearance). **b**, **c**, The sedimentary-basin base elevation (**b**) and thickness (**c**), both derived from gravity modelling. Hairlines in **c** show flight lines. **d**, We interpret regions of differing erosion characteristics for the SSB from the elevation and character of subglacial topography and from variations in the preserved thickness of the sedimentary basins (Extended Data Table 1). Heavy solid and dashed lines show the A/B and B/C boundaries, respectively. Thin lines show the interpreted boundaries of the dendritic erosion pattern for regions C and B1. Thin dotted lines indicate the Cape Goodenough Ridge. The inset shows the location of the region within Antarctica. MUIS, Moscow University Ice Shelf; VF, Vanderford Glacier.

highlands generally either are covered with slow-moving ice, or are ice free, with little scope for fast-moving ice (Fig. 3).

Region A surrounds the coastal glaciers, and is a broad region of moderate elevation, cut by a number of smaller channels that align with modern ice-sheet flow[12]. The SSSB thickness is nil for much of region A and rarely exceeds 1 kilometre (Fig. 1c). The A/B boundary is located near a topographic ridge that extends west–southwest from Cape Goodenough. This ridge averages about −200 metres of elevation, but is cut by channels as deep as −800 metres. The ice-sheet bed within region A is ocean sloping, with an average gradient (measured along flight lines) of +2 to +4 m km^{-1}.

Region B is defined by an almost linear inland-thickening trend to the base of the SSSB, with an upstream limit defined where this thickening trend ceases (Figs 1 and 2). In region B, the thickness of the SSSB increases in co-variance with the reduction in surface ice-sheet velocity (Fig. 2 and Extended Data Figs 3–6). Non-selective and inland-reducing erosion, coupled with the correlation with present-day surface velocity, points to prolonged activity of the modern-scale ice sheet.

Under present climate conditions, ice-sheet modelling indicates high basal velocities within region A (Extended Data Fig. 7a). With an air temperature anomaly, dT_a, of +4 °C and an ocean temperature anomaly, dT_o, of +1 °C (anomaly is with respect to present-day temperatures), a retreat of the ice-sheet margin to the A/B boundary is indicated, with high basal velocities occurring mostly within region B (Fig. 3a). The results of these models suggest that repeated small-scale retreat and advance cycles within region A are necessary to explain the observed erosion. These cycles had a length scale of less than 200 kilometres, did not involve collapse, and were probably orbitally forced[2].

Region B comprises two subregions. Subregion B2 preserves only the modern-scale ice-sheet-erosion signature, and the SSSB is typically more than 2 kilometres thick. In subregion B1, the SSSB is typically thinner and includes branched channels in addition to the modern-scale ice-sheet signature (Fig. 1). The ice-sheet bed in subregion B1 is beneath sea level and typically inland sloping, with an average gradient of −1 to −3 m km^{-1}.

Retreat into the SSB from the topographic ridge at the A/B boundary is subject to ocean-driven instabilities. These include the marine ice-sheet instability, moderated by ice-shelf buttressing[20,21], and ice-cliff failure augmented by hydrofracturing[4]. The former mechanism is widely considered to be the main driver of ice-sheet collapse[21]. The latter mechanism may yield rapid and extensive retreat, although further work is required to verify its role in large-scale ice-sheet collapse[4].

Following retreat into the SSB interior, a new ice-sheet margin is established in front of Highlands B and C. Region C (Fig. 1) is characterized by sufficient erosion to expose basement, and an overall dendritic pattern. Patches of thicker sedimentary rocks exist between channels.

© 2016 Macmillan Publishers Limited. All rights reserved

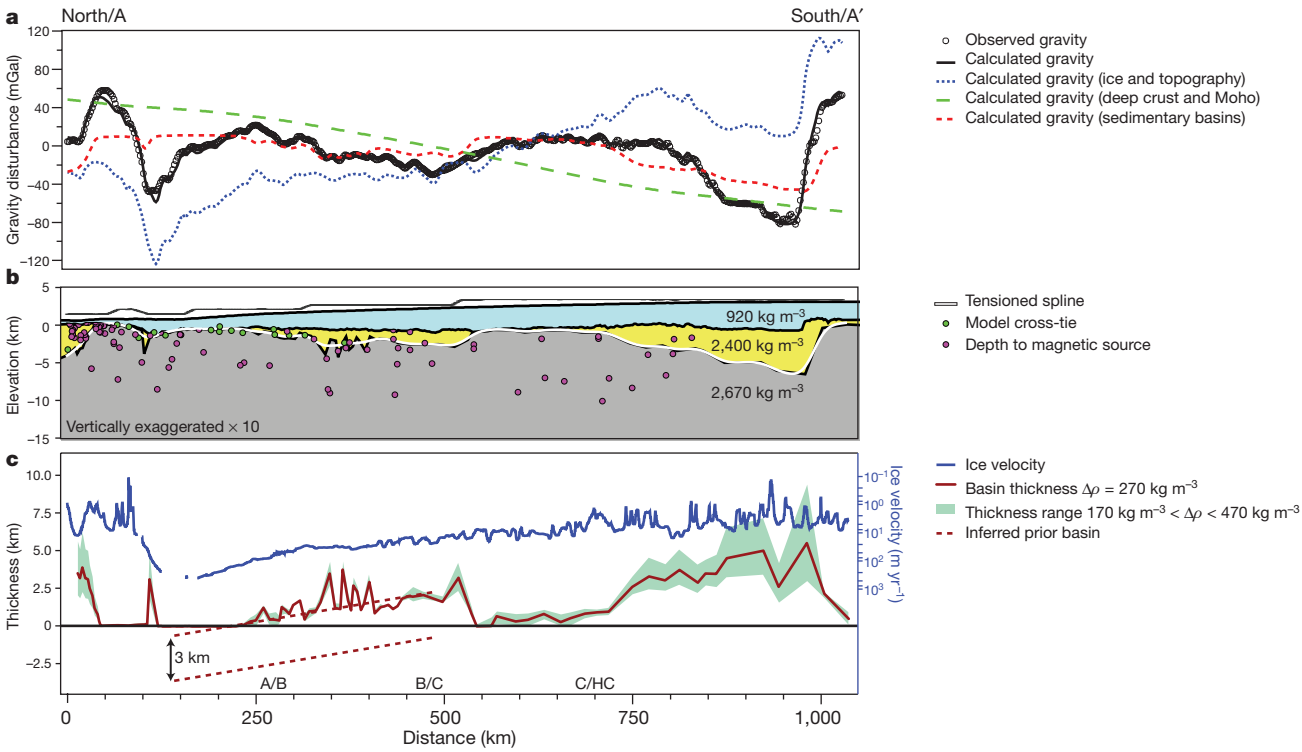


Figure 2 | Gravity model along flight line R06Ea. a, Observed and calculated gravity disturbances, including model components from ice and topography, the deep crust and Moho, and the sedimentary basins. **b**, Results of the model, showing the thickness of sedimentary rocks required to satisfy the gravity field. Estimates of depth to magnetic source[15], the tensioned spline fit, and cross-ties with other models are also shown. The gravity disturbance was modelled at true observation elevations, indicated by the uppermost grey line. **c**, SSSB thickness and present-day surface-ice velocity[12] (vertically flipped; average error in ice-sheet velocity on this line is 10.5 m yr^{-1}). A/B, B/C and C/HC indicate the locations of the interpreted erosion boundaries demarcated in Figs 1 and 3 (HC, Highland C). Dashed lines indicate the inferred prior 3-kilometre thickness of the SSSB, which has been almost entirely removed in region A, and eroded in line with present-day ice velocity in region B. See Extended Data Figs 3–6 for additional representative models throughout the SSB region.

Highland C and the Terre Adelie highlands peak well above sea level, and have average measured gradients on their frontal slopes of about +2.5 m km^{-1}. Consequently, these highlands are persistent barriers to further ice-sheet retreat, even under highly unstable conditions[4,19].

Highland B, however, is pierced by several deep fjords, which indicate extended periods during which fast ice flow was channelled from a large upstream source in the ASB[8]. The fjords are also where ice-sheet retreat into the ASB is probably initiated. This additional vulnerability limits the residence time of an ice margin in front of Highland B as compared with Highland C[4,8]. Our observed erosion in region B1, which is less than the erosion in region C, suggests that retreat of the EAIS into the ASB has occurred as a characteristic of a fully retreated ice sheet.

The erosion characteristics of regions B1 and C point to the activity of a retreated ice sheet. In our models, when $dT_a = +8$ °C and $dT_o = +2$ °C (Fig. 3b), the ice-sheet margin retreats to subregion B1, with high basal velocities mostly within subregion B1 and region C. When $dT_a = +12$ °C and $dT_o = +5$ °C, further retreat to the B/C boundary is suggested (Fig. 3c). This retreat scenario is similar in extent to a previous interpretation[8] of the early ice sheet, before advance into the SSB, although the locations of the ice margins differ. With $dT_a = +15$ °C and $dT_o = +5$ °C, full retreat into the ASB is indicated (Extended Data Fig. 7b). This retreat scenario is somewhat more extensive than a previous interpretation that considered the early ice sheet, before advance into the ASB[8].

Overall, the extent of SSB erosion points to long periods of time during which the SSB was subjected to a modern-scale ice sheet, with the margin located within region A (less than 150 kilometres retreat from present), and also large periods of time when it was subjected to a retreated ice sheet, with the margin located much further inland, near the B/C boundary (350 kilometres of retreat from present). Subregion B1 is apparently less eroded than both regions A and C, suggesting that the residence time of the ice-sheet margin in this region has been less. This may indicate repeated but relatively short-lived transitions between the modern-scale and retreated states.

Our interpretation is not time specific, but global climate and sea-level data suggest that the retreated ice-sheet was predominant in the Oligocene to mid-Miocene, with the modern-scale ice-sheet predominant since the mid-Miocene[22]. Corresponding detrital provenance records at Prydz Bay[1] and locally in sediment cores[23] suggest that basement rocks were being eroded from region A at 7 Ma and at 3.5 Ma (ref. 1). Therefore, the modern-scale ice-sheet has been a recurrent feature of Totten Glacier since well before 7 Ma.

It has been suggested that a substantial Pliocene retreat here may have been necessary to generate the estimated high sea level of up to 22 ± 10 metres above present[24], although a recent estimate is more moderate, at 9–13.5 metres; ref. 25. Our data cannot directly constrain the extent of the Pliocene ice sheet at Totten Glacier, but we can estimate the contribution of this sector to global sea level for our representative ice-sheet models.

Prior reconstructions of retreat from a modern-scale configuration under Pliocene conditions are highly variable[26]. Models with a highly unstable ice sheet[4,26,27] can retreat markedly, while more stable ice-sheet reconstructions typically fail to do so[26,28]. Furthermore, different ice-sheet models can resolve similar ice-sheet extents with differing amounts of sea-level contribution[4,26]. There might also be short-term sea-level fluctuations that we do not consider here[29].

The largest retreat possible under the modern-scale ice-sheet configuration (Fig. 3a) is associated with a total Antarctic sea-level contribution of 8.39 metres, of which the SSB/ASB sector provides 0.89 metres. Even allowing for 7.3 metres of sea-level increase from collapse of the Greenland ice sheet[30], this is insufficient to explain 22 metres of total

© 2016 Macmillan Publishers Limited. All rights reserved

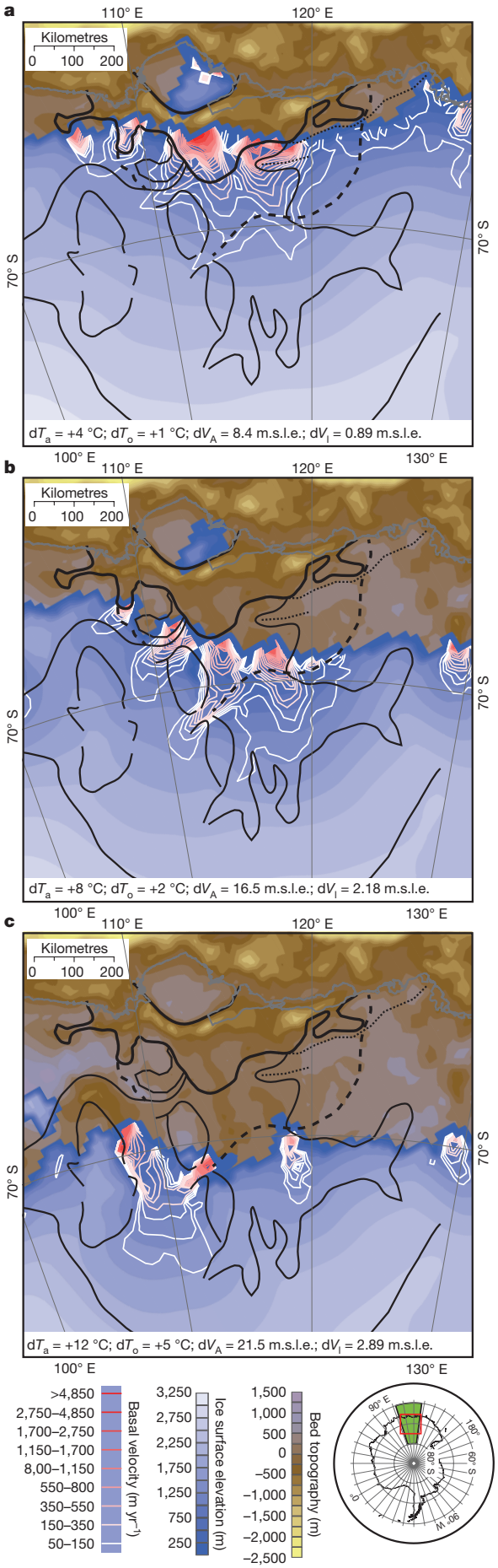


Figure 3 | Ice-sheet models with differing climate forcings. Ice-sheet surface and bed elevations and basal velocity contours are shown after a long-term 20,000-year run under constant climate forcing. Models were run with air temperatures (dT_a) and ocean temperatures (dT_o) above today's. Contributions to global-mean sea level are estimated for all Antarctica (dV_A) and for the SSB/ASB sector (dV_I; indicated by the green area in the inset). **a**, With $dT_a = +4$ °C and $dT_o = +1$ °C, the ice-sheet margin is located near the A/B boundary and high basal velocities are focused in region B. **b**, With $dT_a = +8$ °C and $dT_o = +2$ °C, the ice-sheet margin is located in region B and high basal velocities are focused in regions B1 and C. **c**, With $dT_a = +12$ °C and $dT_o = +5$ °C, the ice-sheet margin is located near the B/C boundary and high basal velocities are focused in region C and the ASB. Further models are shown in Extended Data Fig. 7. m.s.l.e., metres of sea-level equivalent.

sea-level rise[24], but it is sufficient to explain an increase of 13.5 metres (ref. 25).

Our two models of the retreated ice sheet state (Fig. 3b, c) are associated with a total Antarctic sea-level contribution of 16.5 metres or 21.5 metres, of which the SSB/ASB sector provides 2.18 metres and 2.89 metres, respectively. Retreat of the ice sheet into the ASB (Extended Data Fig. 7b) is associated with a total Antarctic sea-level contribution of 29.1 metres, with 4.29 metres sourced from the SSB/ASB sector.

The influence of Totten Glacier on past sea level is clearly notable, but for any particular warm period it is also highly uncertain, because the system is subject to progressive instability. Our results suggest that the first discriminant is the development of sufficient retreat to breach the A/B-boundary ridge. This causes an instability-driven transition from the modern-scale configuration to the retreated configuration. Under ongoing ice-sheet loss, the breaching of Highland B causes further retreat into the ASB. Each of these changes in state is associated with a substantial increase in both the absolute and the proportional contribution of this sector to global sea level.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 9 September 2015; accepted 16 February 2016.

1. Williams, T. *et al.* Evidence for iceberg armadas from East Antarctica in the Southern Ocean during the late Miocene and early Pliocene. *Earth Planet. Sci. Lett.* **290,** 351–361 (2010).
2. Patterson, M. O. *et al.* Orbital forcing of the East Antarctic ice sheet during the Pliocene and Early Pleistocene. *Nature Geosci.* **7,** 841–847 (2014).
3. Cook, C. P. *et al.* Dynamic behaviour of the East Antarctic ice sheet during Pliocene warmth. *Nature Geosci.* **6,** 765–769 (2013).
4. Pollard, D., DeConto, R. M. & Alley, R. B. Potential Antarctic Ice Sheet retreat driven by hydrofracturing and ice cliff failure. *Earth Planet. Sci. Lett.* **412,** 112–121 (2015).
5. Pritchard, H. D., Arthern, R. J., Vaughan, D. G. & Edwards, L. A. Extensive dynamic thinning on the margins of the Greenland and Antarctic ice sheets. *Nature* **461,** 971–975 (2009).
6. Rignot, E., Jacobs, S., Mouginot, J. & Scheuchl, B. Ice shelf melting around Antarctica. *Science* **341,** 266–270 (2013).
7. Fretwell, P. *et al.* Bedmap2: improved ice bed, surface and thickness datasets for Antarctica. *Cryosphere* **7,** 375–393 (2013).
8. Young, D. A. *et al.* A dynamic early East Antarctic Ice Sheet suggested by ice-covered fjord landscapes. *Nature* **474,** 72–75 (2011).
9. Gwyther, D. E., Galton-Fenzi, B. K., Hunter, J. R. & Roberts, J. L. Simulated melt rates for the Totten and Dalton ice shelves. *Ocean Science* **10,** 267–279 (2014).
10. Khazendar, A. *et al.* Observed thinning of Totten Glacier is linked to coastal polynya variability. *Nature Commun.* **4,** 2857 (2013).
11. Greenbaum, J. S. *et al.* Ocean access to a cavity beneath Totten Glacier in East Antarctica. *Nature Geosci.* **8,** 294–298 (2015).
12. Rignot, E., Mouginot, J. & Scheuchl, B. Ice flow of the Antarctic ice sheet. *Science* **333,** 1427–1430 (2011).
13. Wright, A. P. *et al.* Evidence of a hydrological connection between the ice divide and ice sheet margin in the Aurora Subglacial Basin, East Antarctica. *J. Geophys. Res. Earth Surface* **117,** F01033 (2012).
14. Roberts, J. L. *et al.* Refined broad-scale sub-glacial morphology of Aurora Subglacial Basin, East Antarctica derived by an ice-dynamics-based interpolation scheme. *Cryosphere* **5,** 551–560 (2011).
15. Aitken, A. R. A. *et al.* The subglacial geology of Wilkes Land, East Antarctica. *Geophys. Res. Lett.* **41,** 2390–2400 (2014).

© 2016 Macmillan Publishers Limited. All rights reserved

16. Aitken, A. R. A. *et al.* The Australo-Antarctic Columbia to Gondwana transition. *Gondwana Res.* **29,** 136–152 (2016).
17. Drewry, D. J. *Glacial Geologic Processes* (Edward Arnold, 1986).
18. Rose, K. C. *et al.* Early East Antarctic Ice Sheet growth recorded in the landscape of the Gamburtsev Subglacial Mountains. *Earth Planet. Sci. Lett.* **375,** 1–12 (2013).
19. Jamieson, S. S. R., Sugden, D. E. & Hulton, N. R. J. The evolution of the subglacial landscape of Antarctica. *Earth Planet. Sci. Lett.* **293,** 1–27 (2010).
20. Gudmundsson, G. H. Ice-shelf buttressing and the stability of marine ice sheets. *Cryosphere* **7,** 647–655 (2013).
21. Joughin, I. & Alley, R. B. Stability of the West Antarctic ice sheet in a warming world. *Nature Geosci.* **4,** 506–513 (2011).
22. Flower, B. P. & Kennett, J. P. The middle Miocene climatic transition: East Antarctic ice sheet development, deep ocean circulation and global carbon cycling. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **108,** 537–555 (1994).
23. Pierce, E. L. *et al.* Characterizing the sediment provenance of East Antarctica's weak underbelly: the Aurora and Wilkes sub-glacial basins. *Paleoceanography* **26,** PA4217 (2011).
24. Miller, K. G. *et al.* High tide of the warm Pliocene: implications of global sea level for Antarctic deglaciation. *Geology* **40,** 407–410 (2012).
25. Winnick, M. J. & Caves, J. K. Oxygen isotope mass-balance constraints on Pliocene sea level and East Antarctic Ice Sheet stability. *Geology* **43,** 879–882 (2015).
26. de Boer, B. *et al.* Simulating the Antarctic ice sheet in the late-Pliocene warm period: PLISMIP-ANT, an ice-sheet model intercomparison project. *Cryosphere* **9,** 881–903 (2015).
27. Gasson, E., DeConto, R. & Pollard, D. Antarctic bedrock topography uncertainty and ice sheet stability. *Geophys. Res. Lett.* **42,** 5372–5377 (2015).
28. Pollard, D. & DeConto, R. M. Modelling West Antarctic Ice Sheet growth and collapse through the past five million years. *Nature* **458,** 329–332 (2009).
29. Golledge, N. R. *et al.* The multi-millennial Antarctic commitment to future sea-level rise. *Nature* **526,** 421–425 (2015).
30. Alley, R. B. *et al.* History of the Greenland Ice Sheet: paleoclimatic insights. *Quat. Sci. Rev.* **29,** 1728–1756 (2010).

Acknowledgements Collection of ICECAP data was supported by: National Science Foundation grant PLR-0733025; National Aeronautics and Space Administration grants NNX09AR52G, NNG10HPO6C and NNX11AD33G (Operation Ice Bridge and the American Recovery and Reinvestment Act); Australian Antarctic Division projects 3013 and 4077; National Environment and Research Council grant NE/D003733/1; the Jackson School of Geosciences; the G. Unger Vetlesen Foundation; and the Australian Government's Cooperative Research Centres Programme through the Antarctic Climate & Ecosystems Cooperative Research Centre (ACE CRC). Ice-sheet modelling was funded under contract VUW1203 of the Royal Society of New Zealand's Marsden Fund. We thank S. Jamieson for comments and for supplying model images for comparison. This is the University of Texas Institute of Geophysics contribution 2950.

Author Contributions A.R.A.A. undertook data processing, analysis and modelling and wrote the paper. D.D.B., T.D.v.O. and M.J.S. coordinated and planned the fieldwork and led the ICECAP programme. J.L.R., D.A.Y. and J.S.G. undertook the field program and implemented data collection, quality control and initial data processing. N.R.G. conducted the ice-sheet modelling. All authors contributed significantly to interpretation and discussion of the data and manuscript preparation.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.R.A.A. (alan.aitken@uwa.edu.au).

© 2016 Macmillan Publishers Limited. All rights reserved

METHODS

Data. Aerogeophysical data used include ice-surface, ice-thickness, magnetic and gravity data, all of which were collected through the ICECAP program during Antarctic summer seasons 2009/2010 through to 2012/2013. Instruments were flown aboard the DC-3T aircraft, civil registered as C-GJKB, owned and operated by Kenn Borek Air. Line data for these ICECAP products are available at full resolution through the ICEBRIDGE data portal at the National Snow and Ice Data Center (http://nsidc.org/icebridge/portal/).

The ice-surface elevation was obtained using a Riegl LD90-3800-HiP-LR laser ranging system, combined with a range of GPS receivers and inertial measurement unit (IMU) systems. Spatial resolution is 25 m along track and 1 m across track, with a net error of approximately 12 cm (ref. 31). Ice thickness was determined with the 60 MHz HiCARS ice-penetrating radar system[32]. Crossovers in ice thickness yield an average difference of approximately 33 m (ref. 8); crossover difference scales as a function of basal roughness up to a few hundred metres. This error is due predominantly to uncertainty in the horizontal location of the reflection point, which occurs within a radius of about 500 m.

Magnetic intensity data were measured with a Geometrics G-823A caesium vapour magnetometer housed in an aircraft tailboom[33]. Raw total magnetic intensity data are reduced to magnetic intensity anomalies by correcting for the International Geomagnetic Reference Field, time variations in magnetic field, and spline-based line levelling[15].

For the first three seasons, gravitational acceleration data were obtained with a Bell Aerospace BGM-3, two-axis-stabilized scalar gravity meter[34]. Gravitational acceleration was sampled at 1 Hz. Aircraft vertical and horizontal accelerations were estimated using dual carrier phase GPS solutions, and removed from the observed accelerations before filtering. A symmetrical finite impulse response low-pass filter with half-amplitude frequency point of 0.0054 Hz (185 s) was used to smooth the resulting data. In the 2012/2013 season, the BGM-3 gravity meter was replaced with a three-axis-stabilized Canadian Microgravity GT-1A system[35]. Gravitational acceleration was sampled at an effective sampling rate of 18.75 Hz, and filtered with a filter length of 150 s (0.006667 Hz). Typical aircraft velocity was 90 m s^{-1}, so this results in a minimum resolvable half-wavelength of approximately 8 km for BGM-3 data and 6.75 km for GT-1A data.

The gravitational acceleration data were automatically edited on the basis of horizontal and vertical acceleration thresholds, and corrected for latitude, instrument drift, elevation, and the Eotvös effect. The final result is the disturbance from the global gravity field, or gravity disturbance. This is equivalent to the free-air anomaly.

Gravity modelling approach and model uncertainty. For gravity modelling, we used the gravity disturbance data, with vertical gravitational accelerations calculated at observation elevation. For modelling along flight lines, we used an iterative combination of forward modelling and inversion, using the commercial two-dimensional code GM-sys which is based upon the principles of ref. 36. Full three-dimensional modelling is not appropriate for our survey design, owing to the sparse and irregularly spaced data and variably oriented flight lines.

An initial model, described below, was perturbed using an iterative combination of inversion and manual forward modelling. The boundaries of the sedimentary basin base were inverted for depth using the inbuilt GM-sys inversion module, which uses a damped least-squares optimization[37] to estimate the thickness of the basin that minimizes gravity misfit. Inversion damping is not user specified, and results can be erratic. Therefore, in between inversion iterations, we carried out manual editing involving smoothing of anomalously steep or erratic results, and corrections so as to better match the ice-sheet-bed morphology where the thickness is close to zero. Most results were achieved within two or three iterations.

This approach can successfully be used to model one-dimensional and two-dimensional structures, but not complex three-dimensional structures. Our approach is valid for the ice-sheet surface and the SSSB base, which are very close to one-dimensional, but some topographic features are three dimensional in form. In addition, we only resolve gravity features with widths of more than 6.75–8 km, whereas the ice-sheet bed is resolved at a much finer detail. Our models represent this topography with a data spacing of roughly 2 km. Combining the above leads to inaccurate basin-thickness models near rugged topography—such regions including the edges of Highland A, fjords, and deep channels. These regions are also most likely to possess large errors in the subglacial topography data set. In elevation terms, these errors are on average an order of magnitude less than our gravity modelling error, but in gravity terms they might be important owing to the high density contrast at this interface ($\sim$1,500 kg m^{-3} versus $\sim$300 kg m^{-3}).

We constructed an initial model consisting of five layers: ice sheet, sedimentary basins, upper crust, lower crust and mantle. These layers exist, in order, across the entire model and cannot overlap, although they can become zero thickness. Boundaries for the ice-sheet surface and bed elevations were derived directly from the ICECAP line data, and for the ice sheet we used a density of 920 kg m^{-3}. Density variations from the ice-sheet surface and bed are the most notable component of the gravity disturbance data (Fig. 2 and Extended Data Figs 3–6). The Moho was defined by a flexural model that accounts for ice and topographic loads. An elastic thickness of 25 km was used, although—because of the long-wavelength of the loads involved—little variation in Moho structure is observed between 10 km and 50 km elastic thickness. The lower crust had a density of 2,800 kg m^{-3}, the mantle a density of 3,200 kg m^{-3}. Density variations from these interfaces largely balance out the long-wavelength gravity effects of ice-thickness and topography variations (Fig. 2 and Extended Data Figs 3–6). None of these layer boundaries was a variable in the gravity model.

Our target interface for modelling is the base of the sedimentary basins. We precondition the model using previously published depth-to-magnetic-source estimates from the ICECAP data set[15]. These source points were derived from multiple passes of Werner deconvolution[38] using the levelled magnetic line data. As an initial surface, we infer the base of the sedimentary basin at the uppermost resolved magnetic source, under the condition that the slope of the basin base was not excessive. Where the bed is typically deep, we ignored responses at the ice-sheet bed. Substantial scatter is evident in the depth-to-source estimates, but nonetheless they define the overall geometries of the SSSB and the sedimentary basins in the ASB and VSB[15]. Except for initializing the geometry of this surface, we did not use the depth-to-magnetic-source estimates to constrain the geometry of the returned model surface.

We assumed, in the first instance, a crystalline basement density of 2,670 kg m^{-3}. Sedimentary rock density was varied between a lower limit of 2,200 kg m^{-3} and an upper limit of 2,500 kg m^{-3}, providing an effective density contrast range of 300 kg m^{-3}. The lower limit was based on a reasonable maximum porosity of 25%, given the thickness of ice and the likely age of the rocks; the upper limit was defined by the point at which basin thickness departs considerably from magnetic depth-to-basement estimates[15]. These sedimentary densities provide the largest uncertainty regarding basin thickness and create the ranges of sediment thickness shown in Fig. 2c and Extended Data Figs 3c, 4c, 5c and 6c. In the deeper parts of the SSB, the ASB and VSB, this uncertainty can approach several kilometres. Along-line analyses for our selected profiles give a length-weighted mean uncertainty of $\pm$1,239 m, but the length-weighted mean value for the SSB region is $\pm$609 m. Although large, the error is a systematic uncertainty that varies linearly and predictably with basin thickness, and has no influence on the pattern of basin thickness, only the magnitude. Consequently, our interpreted erosion boundaries are not especially sensitive to this uncertainty, but estimates of the amount of erosion are.

A limitation of the gravity method is that only relative gravity differences are modelled, and results for each line are not directly comparable. To correct for this, we levelled the results to a common baseline. We reduced cross-line differences by applying a constant value to a tensioned spline fit through the basin base on each line, so as to minimize the misfit between that line and all cross-ties. These cross-ties (Extended Data Fig. 8) have a mean cross-tie difference—that is, the total difference in thickness—of 607 m (or $\pm$303.5 m), and a median cross-tie difference of 417 m (or $\pm$208.5 m). The highest values are associated with rugged topography, certain flight lines, and the deepest basin regions. Some lines traverse long sections without cross-ties, in particular at the ends of the lines. It is hard to numerically assess the effect of the lack of tie-points; however, the isolated intersections at the south of the image do not show atypical error, although they are above average (Extended Data Fig. 8). The central SSB provides consistently low cross-line differences (Extended Data Fig. 8). This error encapsulates most of the non-systematic errors in the gravity modelling, including those associated with gravity data accuracy, inversion instabilities, three-dimensional gravity effects, topographic errors, and undersampling of small features. The cross-tie errors suggest that, for a particular density contrast, the sedimentary basin thickness can usually be defined to within $\pm$300 m.

A final consideration is the potential influence of lateral variations in density within the region. We note that the gravity data contain the summed gravity field of all sources. Thus we have the fundamental ambiguity that, for any given lateral density distribution, a basin thickness could be found to satisfy the gravity field, and vice versa. In the absence of firm information on these lateral variations, we have adopted a simple approach to modelling the thickness of sedimentary rocks. Nevertheless, lateral variations in density are undoubtedly present. Lateral variations in sedimentary rock density will occur within the previously defined bounds unless they are highly unusual. Basement density variations may also be important, and can be classed either as intra-block or inter-block variability.

Intra-block variability is hard to predict and to test, and cannot be excluded here, but typically is not a major influence on the large-scale gravity field unless large mafic intrusions exist. The error from internal block variability may be comparable in magnitude, but not in pattern, to the variability in sedimentary basin density. For example, a large, 5-km-thick tabular body with a density contrast of 50 kg m^{-3} is equivalent to a $\pm$500 m uncertainty in basin thickness with density

© 2016 Macmillan Publishers Limited. All rights reserved

contrast of 250 kg m^{-3}. In the case of the SSSB, the observed intra-block gravity variations show correlation with the subglacial topography, which includes fairly well defined glacial landscapes, and, for regions A and B, correlates also with the present ice-sheet velocity. An igneous or metamorphic explanation demands an unlikely coincidence, and so we consider it unlikely that basement density variations are the main source of intra-block gravity variability.

Inter-block density variations are potentially more important. Owing to the size and depth extent of crustal blocks, even small variations between them can be substantial. Extended Data Fig. 1 shows that the SSB region possesses a large gravity disturbance high, whereas the Aurora and Vincennes Subglacial Basin regions possess large gravity disturbance lows. The Knox Coast and Law Dome have relatively high gravity, and the Terre Adelie highlands are mixed in response. Tectonic interpretations based predominantly on magnetic data, but also on gravity data[15], indicate that in this region there are several major tectonic structures—including the Totten Shear Zone, along Totten Glacier; the Aurora Fault and the Indo-Australo-Antarctic Suture, which bound the ASB to the northwest and southeast respectively; and the Frost Fault, which bounds the Terre Adelie highlands. Several unnamed faults subdivide the Mawson Craton[15].

A recent model of Australia's crust[39] returns vertically averaged crustal density anomalies for the comparable region that range from +0.2% to +3.8%, with a mean of +2.0% and a standard deviation (σ) of 0.58%. Applying those 2σ limits with a base density value of 2,670 kg m^{-3}, the overall basement density range is approximately 60 kg m^{-3}. We test the impact of a comparable density range on our basin models by applying reasonable basement density contrasts of ±25 kg m^{-3}. Models along flight lines R03Ea, R06Ea and R08Eb allow comparison with the unperturbed models (Fig. 2 and Extended Data Figs 4, 5). The Y07b and GL0092a lines did not cross sufficient major faults to be included here.

Within these models, each major crustal block is assigned a density of 2,670 kg m^{-3}, 2,645 kg m^{-3} or 2,695 kg m^{-3}. To test our models' robustness, we seek the density configuration within these bounds that leads to the flattest basin morphology (Extended Data Fig. 9). Models were inverted once, with no subsequent smoothing, so the results are slightly more rugged than the initial geometry (Extended Data Fig. 9). Automated adjustment of data shift was applied, and so even blocks without changes to density contrast can have changed basin geometries.

Under this testing, substantial changes are made to the VSB and Terre Adelie regions, and to the thickness variation between these regions and the SSB. For the VSB and ASB regions, and for the Terre Adelie region, it is likely that inter-block basement density variations contribute to the gravity anomaly, although it is not known to what extent. This, combined with the large thickness of the basins, raises uncertainty as to the basin geometry.

For the SSB, which is now underlain by a dense crustal block, substantial changes to thickness are observed, but the pattern of basin thickness remains similar (Extended Data Fig. 9). In particular, the inland slope within regions A and B and the contrast between regions B1 and C are persistent features. The A/B boundary shows no correlation with any major tectonic boundary. On line R08Eb, and to a lesser degree on line R06Ea, the B/C boundary is correlated with a notable basement fault. Such a correlation is consistent with the overdeepening of a fault-related topographic depression during glacial activity, and also with the subsequent influence of fjord formation on EAIS erosion. This link with basement structure does not necessarily imply that the observed thickness variations are tectonic in origin. The same fault does not correlate to a significant thickness variation on line R03Ea.

Ice-sheet modelling. We follow an established methodology in our experiments[29], and adopt the same present-day input climatology and tuning parameters (stress balance and basal resistance prescription). The following aspects of our new experiments differ, however. Our experiments start from a thermally equilibriated ice-sheet geometry and run for 20,000 years (20 kyr), so that the modelled ice-sheet is at (or close to) equilibrium. Climate perturbations are applied as linear changes from present-day conditions, between 2,000 and 3,000 model years. The first 2,000 years allow any transient behaviour associated with model initialization to take place in the absence of environmental perturbations, whereas the subsequent 1,000 years force the ice sheet to evolve slowly to changes in air and ocean temperatures and precipitation. The remaining 17 kyr allow a steady state to be reached. All experiments are run at a spatial resolution of 20 km.

As previously[29], we use a sub-grid-scale basal-traction and driving-stress interpolation scheme to allow realistic grounding-line motion[40]. However, we choose not to implement the sub-grid-scale interpolated ice-shelf basal melt component of this scheme, owing to uncertainties as to its validity and the possible resolution dependence of model results[29,41]. Calving is parameterized using horizontal strain rates and a minimum thickness criterion[42,43].

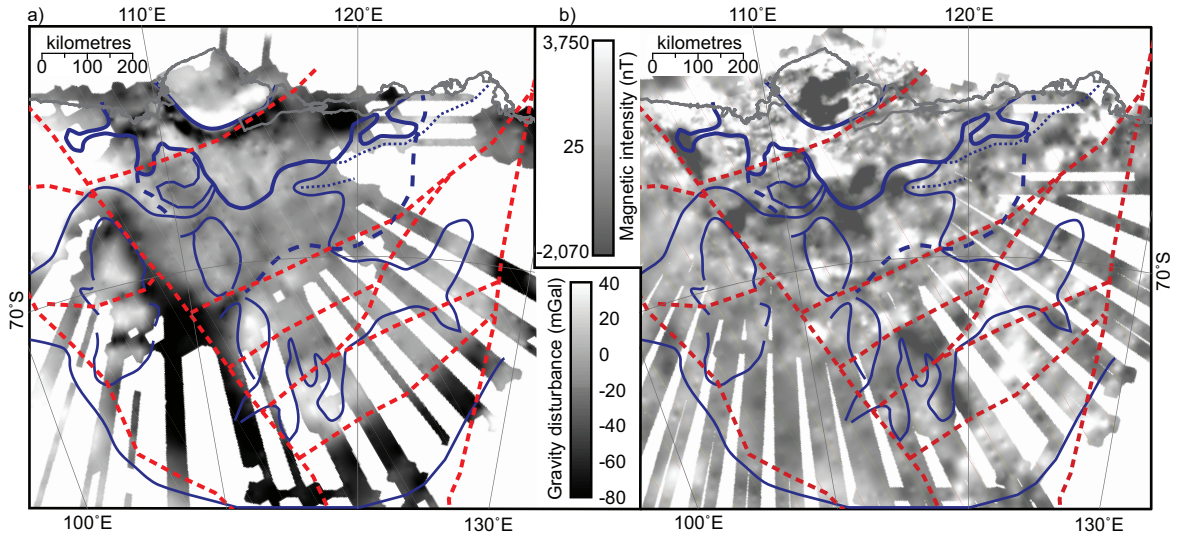
Surface mass balance depends on monthly climatological data and a positive degree-day model that tracks snow thickness and allows for melting of snow and ice at 3 mm °C^{-1} day^{-1} and 8 mm °C^{-1} day^{-1} respectively. We incorporate a white-noise signal (normally distributed; mean random temperature increment of zero) into the calculation of daily temperature variations. The standard deviation of daily temperature variability is set at 2 °C, somewhat lower than the commonly used value of 5 °C, on the basis that the latter has a tendency to overestimate melt[44,45]. Surface temperatures are adjusted for elevation according to an altitudinal lapse rate of −8 K km^{-1}, and a refreezing coefficient of 0.6 is used to mimic meltwater capture within the snowpack.

The sea-level contribution for the whole of Antarctica is derived directly from the numerical ice-sheet model[29]. For the Aurora and Sabrina basins sector, we derive the relative sea-level contribution within the region defined by the longitudinal limits of 102.5° E and 127.5° E, and north of 80° S. These limits encompass the Totten Glacier catchment, with minimal effects from including parts of adjacent regions. We extend the region north into the ocean 60° S so as to capture any isostatic bathymetric effects offshore. For each model result, we compute the remaining mass-loss potential, defined by the difference between the mass of the ice sheet, and the mass of water required to inundate regions below sea level. The difference in mass-loss potential between models returns the mass that has been lost or gained. This mass difference was converted to an equivalent seawater volume, and divided by the global ocean area (3.6×10^{14} m^2) to estimate global sea-level change. For consistency in resolution, we use the model with present-day climate forcing as the baseline, rather than the present-day ice sheet.

Code availability. All geophysical processing and modelling used commercially available software, including ArcGIS from ESRI and Oasis Montaj from Geosoft. Ice-sheet modelling used the PISM code which is available from https://github.com/pism/pism.

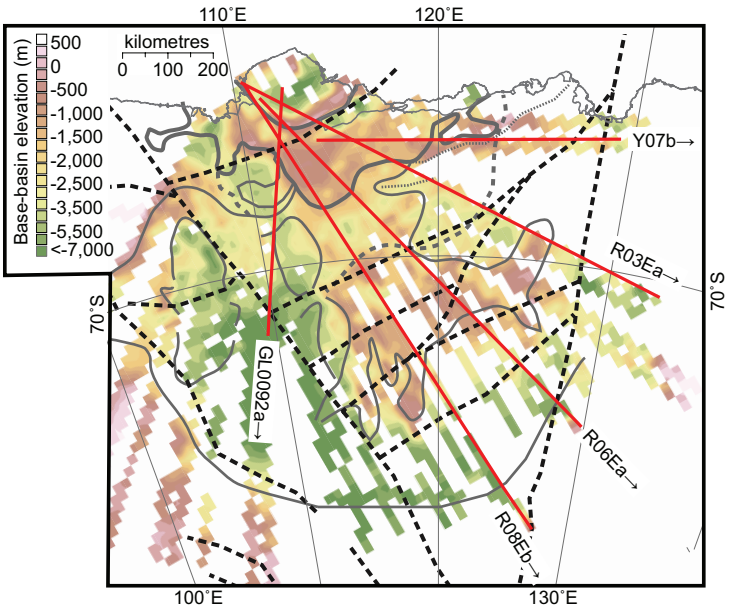
31. Blankenship, D. D. *et al.* IceBridge Riegl Laser Altimeter L2 Geolocated Surface Elevation Triplets, version 1. *NASA Distributed Active Archive Center (DAAC) at the National Snow and Ice Data Center (NSIDC)* http://dx.doi.org/10.5067/JV9DENETK13E (NASA DAAC NSIDC, 2012, updated 2013).
32. Blankenship, D. D., Kempf, S. D. & Young, D. A. IceBridge HiCARS 2 L2 Geolocated Ice Thickness, version 1. *NASA Distributed Active Archive Center (DAAC) at the National Snow and Ice Data Center (NSIDC)* http://dx.doi.org/10.5067/9EBR2T0VXUDG (NASA DAAC NSIDC, 2012, updated 2013).
33. Blankenship, D. D., Kempf, S. D. & Young, D. A. IceBridge Geometrics 823A Cesium Magnetometer L1B Time-Tagged Magnetic Field, version 1. *NASA Distributed Active Archive Center (DAAC) at the National Snow and Ice Data Center (NSIDC)* http://dx.doi.org/10.5067/9WRB1GVWBU7O (NASA DAAC NSIDC, 2011, updated 2013).
34. Blankenship, D. D., Young, D. A., Richter, T. G. & Greenbaum, J. S. IceBridge BGM-3 Gravimeter L2 Geolocated Free Air Anomalies. *NASA Distributed Active Archive Center (DAAC) at the National Snow and Ice Data Center (NSIDC)* http://dx.doi.org/10.5067/8DJW56PKY133 (NASA DAAC NSIDC, 2011, updated 2014).
35. Blankenship, D. D., Young, D. A., Richter, T. G. & Greenbaum, J. S. IceBridge CMG GT-1A Gravimeter L2 Geolocated Free Air Gravity Disturbances. *NASA Distributed Active Archive Center (DAAC) at the National Snow and Ice Data Center (NSIDC)* http://dx.doi.org/10.5067/3X4CIKKSYQRU (NASA DAAC NSIDC, 2014).
36. Talwani, M. Computation with the help of a digital computer of magnetic anomalies caused by bodies of arbitrary shape. *Geophysics* **30,** 797–817 (1965).
37. Webring, M. *SAKI: a Fortran program for generalized linear inversion of gravity and magnetic profiles*. US Geol. Surv. Open-File Report 85-122, http://pubs.usgs.gov/of/1985/0122/report.pdf (1985).
38. Ku, C. C. & Sharp, J. A. Werner deconvolution for automated magnetic interpretation and its refinement using Marquardt's inverse modeling. *Geophysics* **48,** 754–774 (1983).
39. Aitken, A. R. A., Salmon, M. L. & Kennett, B. L. N. Australia's Moho: a test of the usefulness of gravity modelling for the determination of Moho depth. *Tectonophysics* **609,** 468–479 (2013).
40. Feldmann, J., Albrecht, T., Khroulev, C., Pattyn, F. & Levermann, A. Resolution-dependent performance of grounding line motion in a shallow model compared with a full-Stokes model according to the MISMIP3d intercomparison. *J. Glaciol.* **60,** 353–360 (2014).
41. Martin, M. A., Levermann, A. & Winkelmann, R. Comparing ice discharge through West Antarctic Gateways: Weddell vs. Amundsen Sea warming. *Cryosphere Discuss.* **9,** 1705–1733 (2015).
42. Albrecht, T. & Levermann, A. Fracture field for large-scale ice dynamics. *J. Glaciol.* **58,** 165–176 (2012).
43. Levermann, A. *et al.* Kinematic first-order calving law implies potential for abrupt ice-shelf retreat. *Cryosphere* **6,** 273–286 (2012).
44. Seguinot, J. Spatial and seasonal effects of temperature variability in a positive degree-day glacier surface mass-balance model. *J. Glaciol.* **59,** 1202–1204 (2013).
45. Rogozhina, I. & Rau, D. Vital role of daily temperature variability in surface mass balance parameterizations of the Greenland ice sheet. *Cryosphere* **8,** 575–585 (2014).

© 2016 Macmillan Publishers Limited. All rights reserved



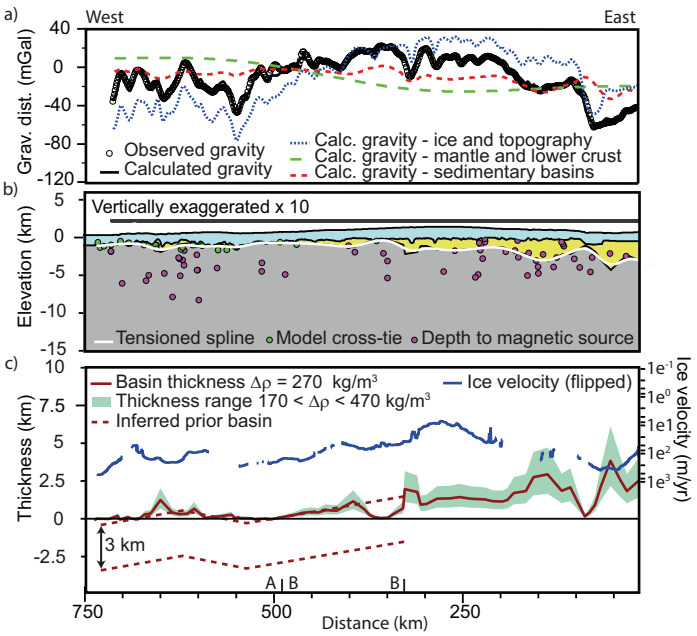
Extended Data Figure 1 | ICECAP geophysical data. a, Gravity disturbance data. **b**, Magnetic intensity data. Dashed red lines indicate interpreted basement faults[15]. Solid, dashed and dotted blue lines indicate the erosion interpretation as shown in Fig. 1.

© 2016 Macmillan Publishers Limited. All rights reserved



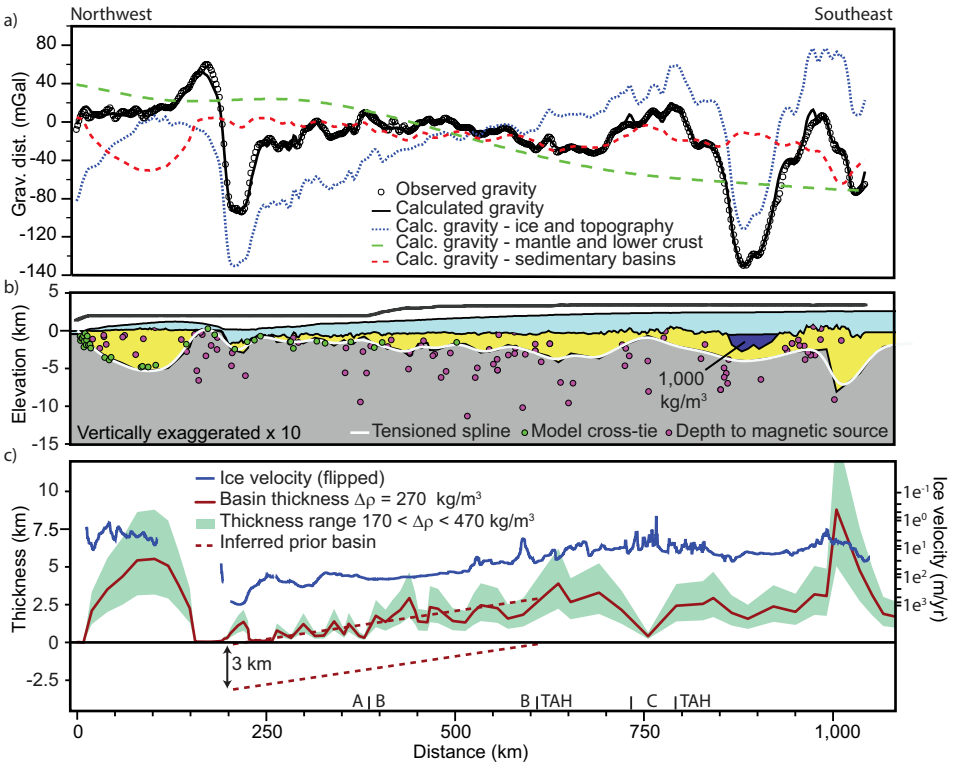
Extended Data Figure 2 | SSSB base morphology. The map shows the relationship of the SSSB base morphology with basement faults (dashed black lines), and the locations of the representative model profiles: flight lines Y07b (Extended Data Fig. 3), R03Ea (Extended Data Fig. 4), R06Ea (Fig. 2), R08Eb (Extended Data Fig. 5) and GL0092a (Extended Data Fig. 6). Arrows indicate the right of the associated model figure in each case. The interpreted erosion boundaries are also shown, in grey, with solid, dashed and dotted lines as in Fig. 1.

© 2016 Macmillan Publishers Limited. All rights reserved



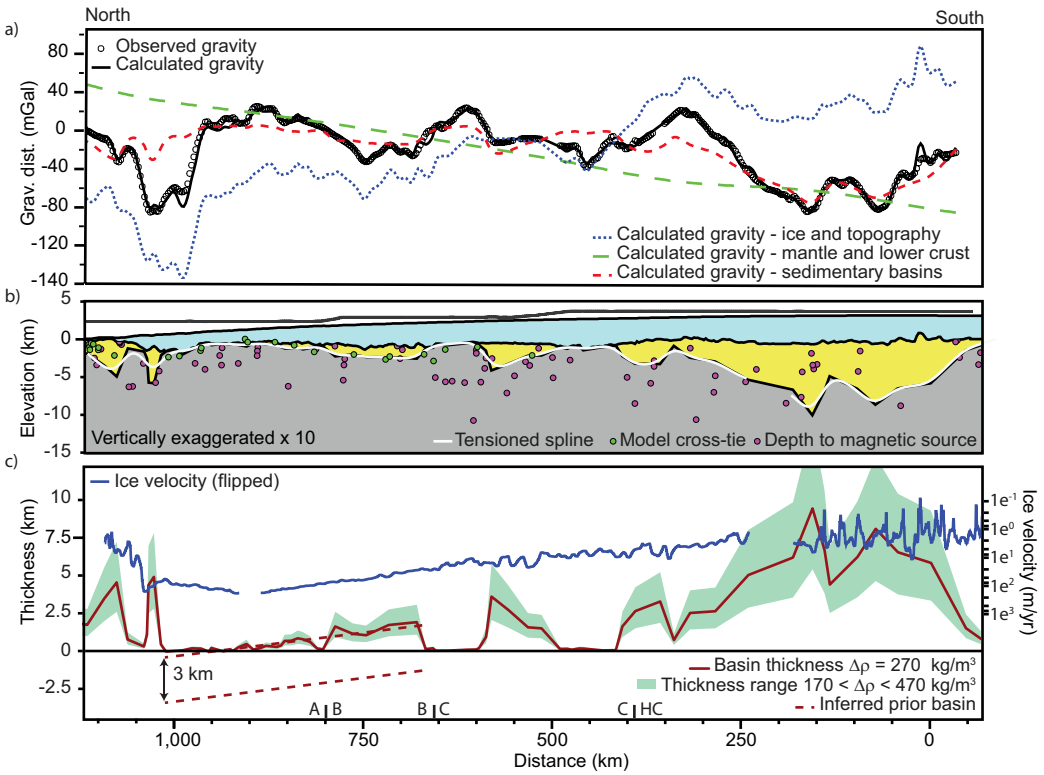
Extended Data Figure 3 | Gravity model along flight line Y07b. a, The observed and calculated gravity disturbance, including model components from ice and topography, the deep crust and Moho ('mantle and lower crust'), and the sedimentary basins. **b**, The model, also showing estimates of depth to magnetic basement, the tensioned spline fit, and cross-ties with other models. Ice, sedimentary rock and basement densities are as per Fig. 2. **c**, SSSB thickness and present-day surface ice velocity (vertically flipped). Basin thickness is shown for a density contrast of 270 kg m^{-3} (2,400 kg m^{-3}), and a thickness range for density contrast of 170–470 kg m^{-3} (2,200–2,500 kg m^{-3}). Average error in ice-sheet velocity on this line is 8.4 m yr^{-1} (ref. 12).

© 2016 Macmillan Publishers Limited. All rights reserved



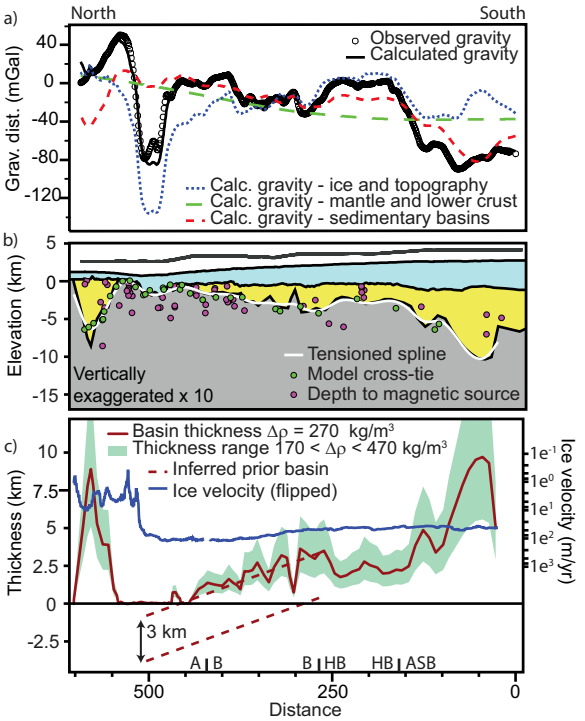
Extended Data Figure 4 | Gravity model along flight line R03Ea. **a**, The observed and calculated gravity disturbance, including model components from ice and topography, the deep crust and Moho ('mantle and lower crust'), and the sedimentary basins. **b**, The model, also showing estimates of depth to magnetic basement, the tensioned spline fit, and cross-ties with other models. Ice, sedimentary rock and basement densities are as per Fig. 2. **c**, SSSB thickness and present-day surface ice velocity (vertically flipped). Basin thickness is shown for a density contrast of 270 kg m^{-3} (2,400 kg m^{-3}), and a thickness range for density contrast of 170–470 kg m^{-3} (2,200–2,500 kg m^{-3}). Average error in ice-sheet velocity on this line is 9.7 m yr^{-1} (ref. 12). TAH, Terre Adelie highlands.

© 2016 Macmillan Publishers Limited. All rights reserved



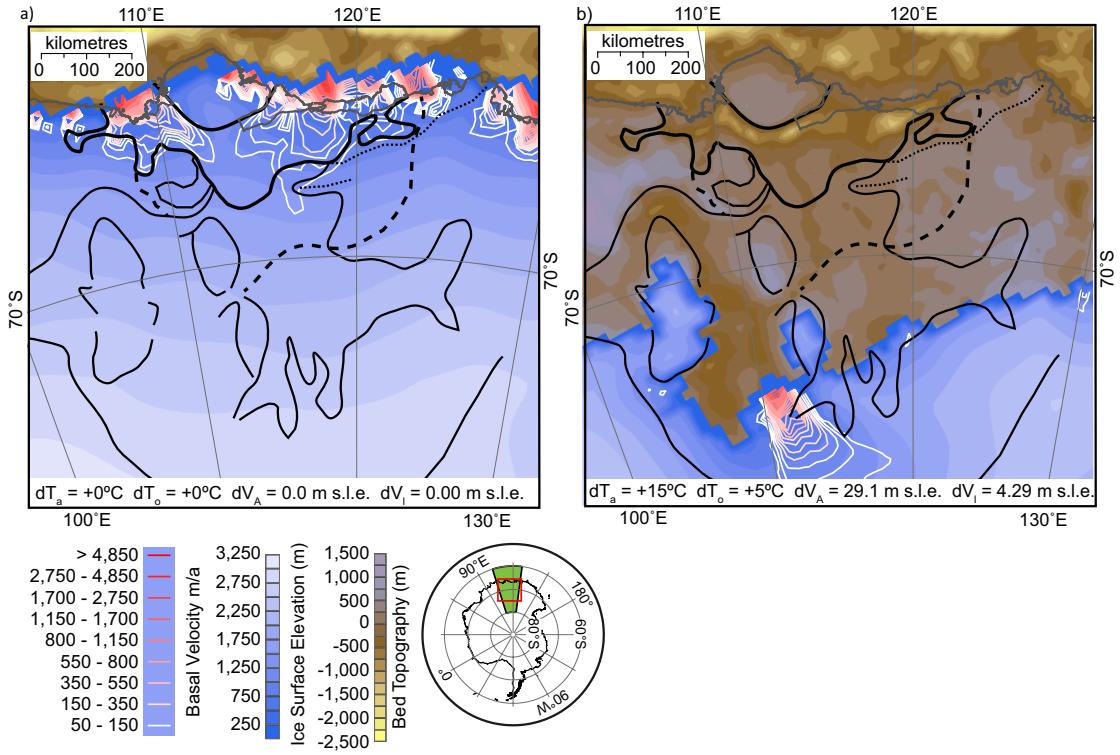
Extended Data Figure 5 | Gravity model along flight line R08Eb. a, The observed and calculated gravity disturbance, including model components from ice and topography, the deep crust and Moho ('mantle and lower crust'), and the sedimentary basins. **b**, The model, also showing depth to magnetic basement estimates, the tensioned spline fit, and cross-ties with other models. Ice, sedimentary rock and basement densities are as per Fig. 2. **c**, SSSB thickness and current surface ice velocity (vertically flipped). Basin thickness is shown for a density-contrast of 270 kg m^{-3} (2,400 kg m^{-3}), and the thickness range for density contrast of 170 to 470 kg m^{-3} (2,200 to 2,500 kg m^{-3}). Average error in ice-sheet velocity on this line is 8.6 m yr^{-1} (ref. 12). HC, Highland C.

© 2016 Macmillan Publishers Limited. All rights reserved



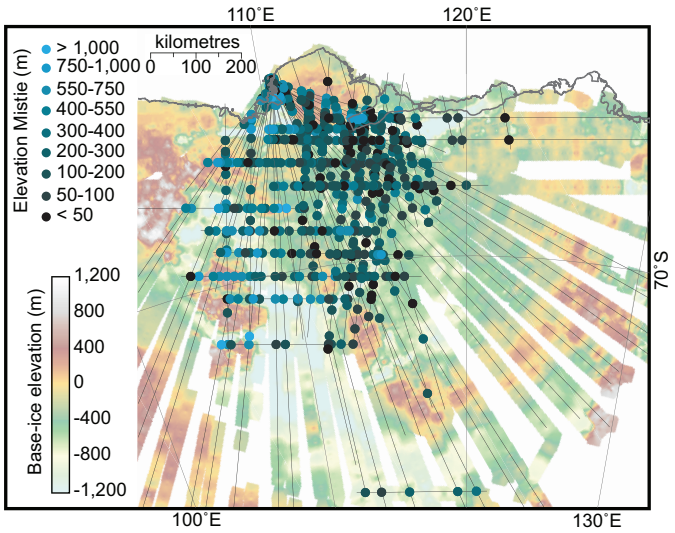
Extended Data Figure 6 | Gravity model along flight line GL0092a. **a**, The observed and calculated gravity disturbance, including model components from ice and topography, the deep crust and Moho ('mantle and lower crust'), and the sedimentary basins. **b**, The model, also showing estimates of depth to magnetic basement, the tensioned spline fit, and cross-ties with other models. Ice, sedimentary rock and basement densities are as per Fig. 2. **c**, SSSB thickness and present-day surface ice velocity (vertically flipped). Basin thickness is shown for a density contrast of 270 kg m^{-3} (2,400 kg m^{-3}), and a thickness range for density contrast of 170–470 kg m^{-3} (2,200–2,500 kg m^{-3}). Average error in ice-sheet velocity on this line is 10.6 m yr^{-1} (ref. 12). HB, Highland B; ASB, Aurora Subglacial Basin.

© 2016 Macmillan Publishers Limited. All rights reserved



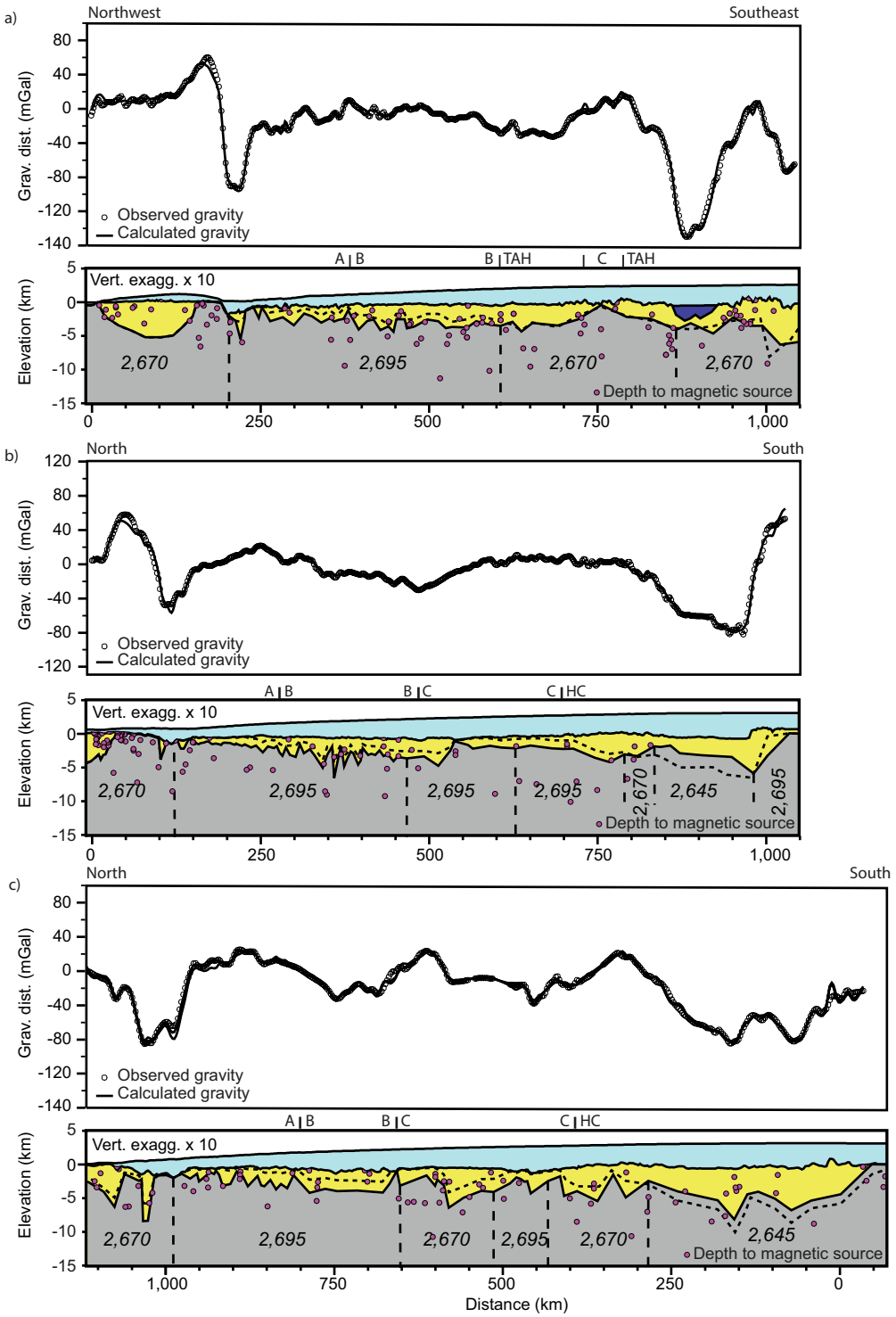
Extended Data Figure 7 | Additional ice-sheet models. Ice-sheet surface and bed elevations and basal velocity contours are shown after a long-term, 20-kyr run under constant climate forcing. **a**, With air and ocean temperatures identical to today's ($dT_a = 0\,°C$ and $dT_o = 0\,°C$), the ice-sheet margin is located near its present location, and high basal velocities are focused in region A. This model includes ice shelves, which also have high basal velocity. **b**, With $dT_a = +15\,°C$ and $dT_o = +5\,°C$, the ice-sheet margin has retreated deep into the ASB and is preserved mainly on inland highlands. Sea-level contributions are defined for all Antarctica (dV_A) and for the ASB/SSB sector (dV_I), denoted by the green region in the inset.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 8 | Model intersection misfits for the gravity models. The dots show the value in metres of each tensioned spline mis-tie after model levelling was applied. Levelling applied a constant value to the base of the sedimentary basin on each line to achieve an optimal minimum-norm fit to all cross-ties. Rugged subglacial topography (background) is the main source of non-systematic model errors because of three-dimensional effects.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 9 | Gravity models including basement density variations. a, Line R03Ea; **b**, line R06Ea; **c**, line R08Eb. Each top panel shows the observed and calculated gravity disturbances. Each bottom panel shows the revised model incorporating basement density contrasts between major crustal blocks. The vertical long-dashed lines indicate the major faults. The dotted line indicates the basin structure with a homogenous basement density of 2,670 kg m^{-3}, as shown in previous figures. Fault locations are derived from the interpretation shown in Extended Data Fig. 1. The fault dip was not tested, as it is a minor component of the field. Blocks are changed by ±25 kg m^{-3} in accordance with models of the comparable region in Australia. Italicized numbers indicate the basement block densities that generated the flattest basin base. Thickness differences are substantial in places, but the overall pattern of basin thickness is preserved. A/B, B/C and so on indicate the interpreted erosional regions as defined in previous figures.

© 2016 Macmillan Publishers Limited. All rights reserved

Extended Data Table 1 | Zonal characteristics of topography and sedimentary rock thickness

Region	Subglacial topography	Sedimentary rock thickness	Bed roughness and basal reflectivity[13]
SSB region A	Deep troughs, reaching -2400m elevation associated with major glaciers (Totten, Moscow University, Vanderford). Hinterland region of generally elevated (ca. -200m) terrain, with a hilly to lineated texture, cut by several channels with up to 600m relief. These channels are aligned with present-day ice flux.	Zero to ~ 1 km, Thinnest at the head of Totten Glacier Thickens towards South-east Thick beneath floating ice shelves	Moderate to high bed roughness High basal reflectivity
SSB region B1	Lower elevation region (ca. -500 m) cut by two ~40 km wide major channels. The main region has a smooth texture. Western channel leads to the ASB via Highland B, and deepens inland from -800m to -1000m in elevation. Eastern channel at ca. -800m elevation leads to region C. This channel possesses several branches leading to the east	Ca. 2-3 km in centre of the zone. Less where channels present, ca. 1-1.5 km, and significantly thinner at zones of confluence (<500m)	Moderate to low bed roughness Low basal reflectivity
SSB region B2	More elevated regions (-400 m to 0m), sloping towards the centre of the SSB. Moderately hilly texture with no major channels.	Ca 1.5-2 km in northwest, 2.5 -3 km in southeast >4 km beneath Highland B	Moderate to high bed roughness, especially at short wavelengths Moderate basal reflectivity
SSB region C	Moderate to very low elevation (0 to -1200 m). Strongly sloping towards the centre of region C. Several branches lead up to Highland C and Terre Adelie highlands, with an overall dendritic pattern.	Typically <500 m. with patchy thicker areas (1-2 km) between channels. One large 3km thick area located between channels at boundary with region B1.	Moderate to high bed roughness Low basal reflectivity
ASB	Very low elevation (-1000 to -1500 m). Smooth texture. Central ridge (-700 to 0 m) separates two deep channels.	Very thick (> 5 km) except in north (2-3 km), and generally thickens to the south	Very low bed roughness High basal reflectivity
VSB	Low elevation (-400 to -800 m), fairly smooth texture. No major channels.	Thick (3-6 km) and generally thickens to the southwest	Low to very low bed roughness High basal reflectivity
Highland Regions	Elevated (>-100m elevation) and typically rough textures. Cut by channels and fjords near ASB.	Low or zero thickness on Ridge B, Dome C, Knox Coast. Moderate to thick thickness on Highlands A, B, C, Law Dome and Terre Adelie Highlands	Typically high bed roughness Typically low basal reflectivity

The listed variations in the characteristics of subglacial topography and the preserved thickness of sedimentary rocks were used to define the zones identified in Fig. 1. The table also documents basal roughness and basal reflectivity characteristics from prior studies[13], although these were not used to guide our interpretation.

© 2016 Macmillan Publishers Limited. All rights reserved

LETTER

doi:10.1038/nature17654

Metabolic acceleration and the evolution of human brain size and life history

Herman Pontzer[1,2], Mary H. Brown[3], David A. Raichlen[4], Holly Dunsworth[5], Brian Hare[6], Kara Walker[6], Amy Luke[7], Lara R. Dugas[7], Ramon Durazo-Arvizu[7], Dale Schoeller[8], Jacob Plange-Rhule[9], Pascal Bovet[10,11], Terrence E. Forrester[12], Estelle V. Lambert[13], Melissa Emery Thompson[14], Robert W. Shumaker[15,16,17] & Stephen R. Ross[3]

Humans are distinguished from the other living apes in having larger brains and an unusual life history that combines high reproductive output with slow childhood growth and exceptional longevity[1]. This suite of derived traits suggests major changes in energy expenditure and allocation in the human lineage, but direct measures of human and ape metabolism are needed to compare evolved energy strategies among hominoids. Here we used doubly labelled water measurements of total energy expenditure (TEE; kcal day^{-1}) in humans, chimpanzees, bonobos, gorillas and orangutans to test the hypothesis that the human lineage has experienced an acceleration in metabolic rate, providing energy for larger brains and faster reproduction without sacrificing maintenance and longevity. In multivariate regressions including body size and physical activity, human TEE exceeded that of chimpanzees and bonobos, gorillas and orangutans by approximately 400, 635 and 820 kcal day^{-1}, respectively, readily accommodating the cost of humans' greater brain size and reproductive output. Much of the increase in TEE is attributable to humans' greater basal metabolic rate (kcal day^{-1}), indicating increased organ metabolic activity. Humans also had the greatest body fat percentage. An increased metabolic rate, along with changes in energy allocation, was crucial in the evolution of human brain size and life history.

Variation in life history reflects evolved differences in energy expenditure. Each organism must allocate its available metabolic energy, which is largely a function of body size, to the competing needs of growth, reproduction and maintenance, resulting in fundamental trade-offs among these vital tasks[2–5]. For example, species that reproduce faster than expected for their body mass generally have shorter maximum lifespans, as energy is directed towards reproductive output and away from maintenance[2–5]. Among primates, this trade-off framework has been expanded to consider the energy needed to grow and maintain large brains[6,7].

In this light, humans present an energetic paradox. Humans in natural fertility populations reproduce more often, and produce larger neonates, than any other living hominoid, yet humans also have the longest lifespans and the largest, most metabolically costly brains[1] (Extended Data Fig. 1). This uniquely human suite of derived, metabolically costly traits suggests a lifting of energetic constraints in the hominin lineage, but the critical underlying mechanisms remain largely unknown. Some have hypothesized that a reduced gut[8] or increased locomotor efficiency[9] provided the extra energy needed for brain expansion. However, phylogenetically informed analyses suggest that gut reduction is insufficient to explain increased human brain size[1], and, although human walking is more economical[10], traditional hunter–gatherers travel so much farther per day[11] that their daily ranging costs are no lower than wild chimpanzees' (Supplementary Discussion and Supplementary Table 1). Similarly, provisioning of young offspring and their mothers helps to shorten human inter-birth intervals and increase the pace of reproduction[1], but reproduction remains relatively costly for human mothers (Supplementary Discussion, Supplementary Fig. 1 and Supplementary Table 2). Furthermore, it is unclear whether food sharing or other dietary changes are sufficient to fuel larger brains, larger neonates and longer lifespans without an acceleration in metabolic rate to harness increased energy acquisition.

Here, we test the hypothesis that humans have evolved an accelerated metabolic rate and larger energy budget, accommodating larger brains, greater reproductive output and longer lifespans without the expected energetic trade-offs. Increased TEE (kcal day^{-1}) has previously been discounted as an explanation for the human energetic paradox[1,8], in part because human basal metabolic rate (BMR; kcal day^{-1}) is broadly similar to that of other primates[8,11]. However, variation in TEE and BMR among humans and the great apes, our closest evolutionary relatives, is largely unstudied. Lacking sufficient data on TEE and BMR in apes, previous analyses have been unable to compare metabolic rates across the entire hominoid clade and test for metabolic acceleration in humans.

We used the doubly labelled water method[12] to measure TEE in mixed-sex samples of adult chimpanzees (*Pan troglodytes*; $n = 27$), bonobos (*Pan paniscus*; $n = 8$), Western lowland gorillas (*Gorilla gorilla*; $n = 10$) and orangutans (*Pongo* spp.; $n = 11$), and compared these data to similar measures of TEE in a large, adult human sample[13] (*Homo sapiens*; $n = 141$); this method also provides a measure of body composition (Methods and Table 1). TEE was measured over 7–10 days while individuals followed their normal daily routine. We also compared published measurements of BMR in humans, chimpanzees and orangutans, and we estimated daily locomotor energy expenditure and BMR for adults in our TEE sample to assess their contribution to variation in TEE. Comparisons were performed at the genus level, both to avoid the issue of close phylogenetic relatedness for chimpanzees and bonobos, and because no metabolic differences were apparent between these two species.

Humans exhibited greater TEE than other hominoids, with larger daily energy budgets than all apes except adult male gorillas

[1]Department of Anthropology, Hunter College. 695 Park Avenue, New York, New York 10065, USA. [2]New York Consortium for Evolutionary Primatology, New York, New York 10065, USA. [3]Lester E. Fisher Center for the Study and Conservation of Apes, Lincoln Park Zoo. Chicago, Illinois 60614, USA. [4]School of Anthropology, University of Arizona, 1099 E South Campus Drive, Tucson, Arizona 85716, USA. [5]Department of Sociology & Anthropology, University of Rhode Island, 45 Upper College Rd, Kingston, Rhode Island 02881, USA. [6]Department of Evolutionary Anthropology, Duke University, Durham, North Carolina 27708, USA. [7]Public Health Sciences, Stritch School of Medicine, Loyola University Chicago, 2160 South First Avenue, Maywood, Illinois 60153, USA. [8]Nutritional Sciences, Biotechnology Center, University of Wisconsin–Madison, 425 Henry Mall, Madison, Wisconsin 53705, USA. [9]Kwame Nkrumah University of Science and Technology, Kumasi, Ghana. [10]Institute of Social & Preventive Medicine, Lausanne University Hospital, Rue de la Corniche 10, 1010 Lausanne, Switzerland. [11]Ministry of Health, PO Box 52, Victoria, Mahé, Seychelles. [12]UWI Solutions for Developing Countries, The University of the West Indies, 25 West Road, UWI Mona Campus, Kingston 7, Jamaica. [13]Research Unit for Exercise Science and Sports Medicine, University of Cape Town, PO Box 115, Newlands 7725, Cape Town, South Africa. [14]Department of Anthropology, University of New Mexico. Albuquerque, New Mexico 87131, USA. [15]Indianapolis Zoo, 1200 W Washington Street, Indianapolis, Indiana 46222, USA. [16]Department of Anthropology and Center for Integrated Study of Animal Behavior, Indiana University, 701 E Kirkwood Avenue, Bloomington, Indiana 47405, USA. [17]Krasnow Institute for Advanced Study, George Mason University, 4400 University Dr., Fairfax, Virginia 22030, USA.

© 2016 Macmillan Publishers Limited. All rights reserved

Table 1 | Age, body size and composition, and TEE

Genus	Sex	*n*	Age (years) (s.d.)	Mass (kg) (s.d.)	FFM (kg) (s.d.)	Body fat* (%) (s.d.)	TEE† (kcal day^{-1}) (s.d.)	BMR‡	Walk and climb‡
Homo	Female	76	33.5 (6.0)	80.1 (19.8)	46.1 (7.4)	41.1 (7.1)	2,190 (381)	1,474	183
	Male	65	32.6 (5.4)	73.8 (15.2)	56.1 (7.9)	22.9 (7.6)	2,721 (569)	1,640	226
Pan	Female	17	25.8 (11.3)	46.4 (8.1)	43.3 (7.8)	9.0 (5.5)	1,722 (363)	1,214	102
	Male	18	22.3 (10.7)	57.9 (13.4)	53.8 (11.7)	8.4 (4.9)	2,145 (546)	1,401	120
Gorilla	Female	6	19.7 (4.1)	73.7 (7.3)	63.3 (7.0)	13.9 (6.4)	2,030 (527)	-	110
	Male	4	24.5 (11.7)	166.4 (42.5)	148.3 (27.0)	15.2 (4.6)	3,630 (839)	-	253
Pongo	Female	5	27.6 (5.3)	58.2 (4.0)	44.3 (2.5)	23.4 (8.8)	1,476 (268)	984	88
	Male	6	20.0 (10.8)	76.7 (34.1)	62.4 (23.3)	16.4 (8.6)	1,617 (289)	1,090	163

s.d., standard deviation.
*Excludes $n = 6$ *Pan* subjects with negative calculated body fats (Methods).
†Calculated for *Pan, Gorilla*, and Pongo using equation 7.17 in ref. 12.
‡Estimated BMR and locomotion costs were calculated from body mass and daily activity (Methods).

(Table 1). Fat-free mass (FFM) explained 34% of the variance in TEE ($t(195) = 10.18$, $r^2 = 0.34$, $P < 0.001$; Fig. 1 and Supplementary Table 3). Differences among genera accounted for an additional 27% of the variance in TEE, with humans having the highest TEE, followed in order by *Pan*, *Gorilla* and *Pongo* ($P \leq 0.001$ for all comparisons; Fig. 1 and Supplementary Table 3). For comparison, mean FFM was similar for *Homo*, *Pan* and *Pongo* cohorts, yet mean human TEE was ~27% greater than *Pan* and ~58% greater than *Pongo*. Body fat was weakly, negatively correlated with TEE ($\beta = -0.04 \pm 0.02$, $P = 0.02$); age ($P = 0.70$) and sex ($P = 0.26$) were not significant covariates in TEE models including FFM, fat mass and genus (Supplementary Table 3). Results were robust across different models for calculating TEE from isotope enrichment (Supplementary Table 3).

Together, BMR and daily locomotor expenditure accounted for much of the variation in TEE among genera. The mean estimated daily energy cost of walking and climbing among apes in our sample was similar across genera and ~100 kcal day^{-1} below estimated daily walking cost in our human sample (Table 1). Our analyses of published measurements[14–16] indicate that BMR is lower in *Pan* than in comparably aged humans, in analyses controlling for sex, age and body mass ($P < 0.001$; Methods and Extended Data Fig. 2). The few measurements of *Pongo* BMR[16,17] appear to be lower still (Extended Data Fig. 2); no data for the *Gorilla* BMR are available. For cohorts in the TEE data set, estimated BMR for *Homo* was ~200 kcal day^{-1} greater than *Pan* and ~500 kcal day^{-1} greater than *Pongo* (Table 1). Estimated physical activity levels (PALs: TEE/estimated BMR; see Table 1), were similar across genera (*Homo*: female 1.5, male 1.7; *Pan*: female 1.4, male 1.5; *Pongo*: female 1.5, male 1.5), indicating that organ activity contributed more than physical activity to variations in TEE.

Body fat percentage, measured from isotope dilution[12], was markedly higher in humans than other hominoids in our TEE sample, and only humans exhibited a significant sex difference (Table 1). Body fat percentage was remarkably low for *Pan*, consistent with measures from dissection[18], and somewhat higher for *Gorilla* and *Pongo*. Indeed, body fat percentages for captive apes were comparable to, or even below, average body fat percentages for humans in physically active, traditional hunter–gatherer populations[19].

Together, these results indicate substantial metabolic evolution across the hominoid clade and a previously uncharacterized diversity in TEE, BMR and body fat percentage (Fig. 2). For a given FFM, humans have greater TEE, BMR and fat mass than other hominoids. Chimpanzees and bonobos have the next largest energy budgets but carry the least amount of body fat. Gorillas have lower TEE and greater body fat than *Pan*. Orangutans have exceptionally low TEE and BMR, and fat percentages similar to gorillas (Table 1).

Comparisons of BMR and TEE indicate that hominin brain and life history evolution were fuelled in large part by increased mass-specific organ metabolic rates (kcal g^{-1} h^{-1}), and not solely through anatomical or physiological trade-offs. BMR reflects the summed metabolic activity of the organs at rest and is largely shaped by organs with high mass-specific metabolic rates such as the brain, liver and gastrointestinal tract[20]. If, as assumed in previous studies, organ-mass-specific metabolic rates were similar across hominoids, then humans and apes with the same FFM should have similar BMRs — the cost of humans' larger brain offset by a reduced gastrointestinal tract[8]. Instead, BMR and TEE are substantially greater for humans, controlling for FFM (Table 1, Figs 1, 2 and Extended Data Fig. 2), indicating an evolutionary increase in the metabolic activity of at least some organ systems. Humans' greater BMR and TEE could readily accommodate the marginal cost of increased reproductive output, estimated at ~130 kcal day^{-1} over the course of a woman's reproductive career in natural fertility populations (Supplementary Discussion, Supplementary Table 2 and Supplementary Fig. 3). Greater metabolic activity could also support humans' derived brain metabolism[21] and increased somatic maintenance and longevity[1], but more comparative analyses are needed to identify specific mechanisms and associated metabolic costs.

Analysing relatively sedentary human and ape populations allowed us to compare cohorts with similar FFMs, PALs and locomotor expenditures (Table 1), but the pattern of TEE, BMR and body composition variation among genera is expected to hold across more physically active populations. When controlling for body size, TEE is remarkably consistent across populations within a species, regardless of activity level[22]: traditional hunter–gatherers have similar size-corrected TEEs to people living in cities[19], and zoo-housed populations of primates and other mammals are similar to their counterparts in the wild[22–24]. Humans in traditional foraging and farming populations and apes in the wild will generally have lower body fat percentages[19] than the cohorts in the TEE sample (Table 1), but again the pattern of differences seen here is expected to persist: even in captivity,

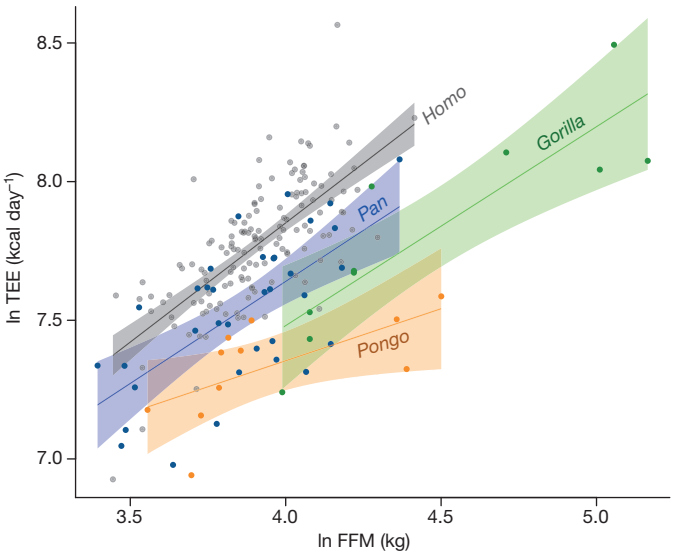


Figure 1 | TTE and FFM for hominoids. Humans (*Homo*, grey, $n = 141$), chimpanzees and bonobos (*Pan*, blue, $n = 35$), gorillas (*Gorilla*, green, $n = 10$), and orangutans (*Pongo*, orange, $n = 11$). Lines and shaded regions indicate least squares regressions and 95% confidence intervals for each genus. TEE for *Homo* exceeds other genera in a general linear model accounting for FFM, fat mass and other variables ($P < 0.001$; Table 1 and Supplementary Table 3).

© 2016 Macmillan Publishers Limited. All rights reserved

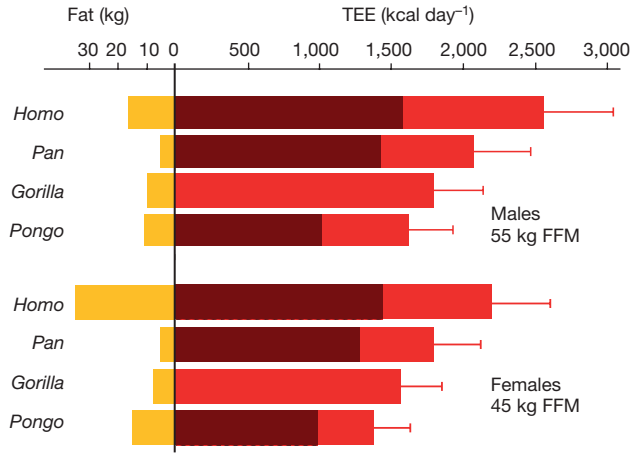


Figure 2 | Predicted TEE, BMR and fat mass for adult hominoids. Values are estimated for males (55 kg FFM) and females (45 kg FFM), using the same FFM across genera. TEE (red bars) is estimated from FFM, fat mass and genus using model C in Supplementary Table 3; error bars represent model standard error. BMR (darker red regions) is estimated from body mass (Methods and Extended Data Fig. 2); no BMR data are available for *Gorilla*. Fat mass (yellow bars) is calculated from FFM using body fat percentages in Table 1.

body fat percentage in *Pan* is remarkably low and fat percentages for *Pongo* and *Gorilla* fall near the very lowest ranges for healthy, physically active human populations[19].

Expansion of the hominin energy budget challenges current models of life history and brain evolution predicated on metabolic stasis and constraint[1,2,5–9]. Evolved increases in BMR and TEE do not diminish the importance of gut reduction, efficient walking, dietary change and maternal provisioning in the hominin lineage, but rather place these critical adaptations within a different energetic framework. Increased TEE exposes individuals to a greater likelihood of energy shortfalls, providing strong selection for behavioural and anatomical adaptations to mitigate this risk. Gut reduction[8] allows for greater energy allocation to reproduction and maintenance (including larger brains) while limiting the increase in BMR and TEE. Improved walking efficiency[10], a dietary shift to more energy-dense foods (for example, meat, tubers), and the adoption of cooking[25] all effectively increase the net energy gained from foraging, and may have had an essential role in the evolutionary expansion of the hominin energy budget. Food sharing[1] provides an additional source of food energy, and along with increased body fat[9], provides an important buffer against energy shortfalls, particularly for children and for females shouldering the added energetic burdens of reproduction. We hypothesize that food sharing and increased body fat coevolved with greater TEE to mitigate the inherent risk of increased energy demands.

Metabolic measurements of living hominoids resolve the human energetic paradox and may also point towards strategies for combating obesity and metabolic disease. Humans exhibit an evolved predisposition to deposit fat whereas other hominoids remain relatively lean, even in captivity where activity levels are modest. Untangling the evolutionary pressures and physiological mechanisms shaping the diversity of metabolic strategies among living hominoids may aid efforts to promote and repair metabolic health for humans in industrialized populations and apes in captivity.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 4 January; accepted 11 March 2016.
Published online 4 May 2016.

1. Isler, K. & van Schaik, C. P. How our ancestors broke through the gray ceiling. *Curr. Anthropol.* **53,** S453–S465 (2012).
2. Charnov, E. L. *Life History Invariants* (Oxford Univ. Press, 1993).
3. Brown, J. H., Gillooly, J. F., Allen, A. P., Savage, V. M. & West, B. G. Toward a metabolic theory of ecology. *Ecology* **85,** 1771–1789 (2004).
4. Stearns, S. C. *The Evolution of Life Histories* (Oxford Univ. Press, 1992).
5. Charnov, E. L. The optimal balance between growth rate and survival in mammals. *Evol. Ecol. Res.* **6,** 307–313 (2004).
6. Charnov, E. L. & Berrigan, D. Why do primates have such long life spans and so few babies? *Evol. Anthropol.* **1,** 191–194 (1993).
7. Isler, K. & van Schaik, C. P. The Expensive Brain: a framework for explaining evolutionary changes in brain size. *J. Hum. Evol.* **57,** 392–400 (2009).
8. Aiello, L. C. & Wheeler, P. The expensive tissue hypothesis. *Curr. Anthropol.* **36,** 199–221 (1995).
9. Navarrete, A., van Schaik, C. P. & Isler, K. Energetics and the evolution of human brain size. *Nature* **480,** 91–93 (2011).
10. Pontzer, H., Raichlen, D. A. & Rodman, P. S. Bipedal and quadrupedal locomotion in chimpanzees. *J. Hum. Evol.* **66,** 64–82 (2014).
11. Leonard, W. R. & Robertson, M. L. Comparative primate energetics and hominid evolution. *Am. J. Phys. Anthropol.* **102,** 265–281 (1997).
12. Speakman, J. R. *Doubly Labelled Water: Theory & Practice* (Chapman & Hall, 1997).
13. Pontzer, H. *et al.* Constrained total energy expenditure and metabolic adaptation to physical activity in adult humans. *Curr. Biol.* **26,** 410–417 (2016).
14. Butte, N. F. Fat intake of children in relation to energy requirements. *Am. J. Clin. Nutr.* **72** (suppl.), 1246S–1252S (2000).
15. Bruhn, J. M. & Benedict, F. G. The respiratory metabolism of the chimpanzee. *Proc. Am. Acad. Arts Sci.* **71,** 259–326 (1936).
16. Bruhn, J. M. The respiratory metabolism of infrahuman primates. *Am. J. Physiol.* **110,** 477–484 (1934).
17. Pontzer, H., Raichlen, D. A., Shumaker, R. W., Ocobock, C. & Wich, S. A. Metabolic adaptation for low energy throughput in orangutans. *Proc. Natl Acad. Sci. USA* **107,** 14048–14052 (2010).
18. Zihlman, A. L. & Bolter, D. R. Body composition in *Pan paniscus* compared with *Homo sapiens* has implications for changes during human evolution. *Proc. Natl Acad. Sci. USA* **112,** 7466–7471 (2015).
19. Pontzer, H. *et al.* Hunter-gatherer energetics and human obesity. *PLoS ONE* **7,** e40503 (2012).
20. Wang, Z., Bosy-Westphal, A., Schautz, B. & Müller, M. Mechanistic model of mass-specific basal metabolic rate: evaluation in healthy young adults. *Int. J. Body Compos. Res.* **9,** 147 (2011).
21. Bauernfeind, A. L. *et al.* Evolutionary divergence of gene and protein expression in the brains of humans and chimpanzees. *Genome Biol. Evol.* **7,** 2276–2288 (2015).
22. Pontzer, H. Constrained total energy expenditure and the evolutionary biology of energy balance. *Exerc. Sport Sci. Rev.* **43,** 110–116 (2015).
23. Pontzer, H. *et al.* Primate energy expenditure and life history. *Proc. Natl Acad. Sci. USA* **111,** 1433–1437 (2014).
24. Nie, Y. *et al.* Exceptionally low daily energy expenditure in the bamboo-eating giant panda. *Science* **349,** 171–174 (2015).
25. Carmody, R. N., Weintraub, G. S. & Wrangham, R. W. Energetic consequences of thermal and nonthermal food processing. *Proc. Natl Acad. Sci. USA* **108,** 19199–19203 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank participating zoos and staff for their efforts: Houston Zoo, Indianapolis Zoo, Jacksonville Zoo, Lincoln Park Zoo, Milwaukee County Zoo, North Carolina Zoo, Oklahoma City Zoo, Oregon Zoo, Zoo Atlanta, Woodland Park Zoo, Dallas Zoo, Brookfield Zoo and Columbus Zoo. We thank B. Moumbaka for assistance administering doses and collecting samples for analysis. We thank R. Atencia and C. Andre for supporting this project. Work at Tchimpounga and Lola Ya Bonobo was performed under the authority of the Ministry of Research and the Ministry of Environment in the Democratic Republic of Congo (research permit #MIN.RS/SG/004/ 2009) and the Ministry of Scientific Research and Technical Innovation in the Congo Republic (research permit 09/MRS/DGRST/ DMAST), with samples imported under CITES permits 09US223466/9 and 9US207589/9. L. Christopher, K. Stafford and J. Paltan assisted with sample analyses. Funding was provided by the US National Science Foundation (BCS-1317170), National Institutes of Health (R01DK080763), L.S.B. Leakey Foundation, Wenner-Gren Foundation (Gr. 8670), University of Arizona and Hunter College.

Author Contributions H.P. and S.R.R. designed the study; H.P., M.H.B., D.A.R., H.D., B.H., K.W., A.L., L.R.D., J.P.-R., P.B., T.E.F., E.V.L., R.W.S. and S.R.R. collected data; H.P., R.D.-A., M.E.T. and D.S. analysed data. All authors contributed to writing the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to H.P. (hpontzer@hunter.cuny.edu).

© 2016 Macmillan Publishers Limited. All rights reserved

METHODS

Sample. Adult apes (age ≥ 10 years) were housed in Association of Zoos and Aquariums accredited zoos in the United States, or in sanctuaries in the Democratic Republic of Congo, with large enclosures. No statistical methods were used to predetermine sample size. Institutional approvals were obtained from Hunter College and participating facilities before data collection. Human data were drawn from a recent large global study[13] (United States, Ghana, South Africa, Seychelles, Jamaica; institutional approvals from local governments and Loyola University, and each subject's informed consent, were obtained before data collection). Men and women employed in manual labour were excluded from analyses to better match activity levels across species. Apes were excluded from analyses if they had a dilution space ratio less than 1.00, or had a calculated body fat percentage below −10%, either of which could signal an error in dosing or sample processing. Owing to normal measurement error, the isotope dilution method for calculating body composition can produce marginally negative body fat estimates in subjects with fat percentages near 0. This occurred in $n = 6$ *Pan* subjects (range: −1 to −6% body fat), which is unsurprising given the low body fat percentage in *Pan*[18] (Table 1). These subjects were included in analyses unless otherwise noted; they were excluded from analyses including fat mass (Supplementary Table 3), because negative values cannot be log-transformed. Pregnant and lactating mothers ($n = 2$ bonobo and 2 gorillas) were retained in the sample because 'pregnant/lactating' did not emerge as a significant factor for TEE in multiple regression with FFM and genus ($t(191) = -0.48$, $P = 0.68$). The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Dosing and sampling. Dosing, sampling and analytical protocols for humans were as described previously[13]. Ape protocols were as described previously[23], and apes from that study ($n = 14$ *Pan*, $n = 3$ *Pongo*, $n = 5$ *Gorilla*) are included in this analysis. Apes drank a sufficient dose of DLW (6% ^2H_2O, 10% $H_2{}^{18}O$; Sigma-Aldrich) to achieve targeted enrichment for their body mass[12], which was measured using a digital scale before the study. Urine or saliva samples were collected once before dosing and then 3–5 times post-dose, over the course of 7–10 days. Isotopic enrichment (^2H and ^{18}O) was measured using isotope ratio mass spectrometry or cavity ring down mass spectrometry. Analysis of residuals indicates results were similar across laboratories and methods (Supplementary Fig. 2).

Calculating TEE. Isotope depletion rates and dilution spaces were calculated via the slope-intercept method[12]. Three to five post-dose urine samples were analysed for each subject, enabling us to identify and remove contaminated samples from analysis. Isotope depletion rates were used to calculate ape TEE (kcal day^{-1}) using equation 7.17 in ref. 12, using a food quotient of 0.95 (ref. 23) and a dilution space ratio of 1.04 following observed ratios in all ape species (1.04 ± 0.02, $n = 56$ subjects; no effect of genus: $F(53) = 0.55$, $P = 0.58$ ANOVA). We also calculated TEE using equation 17.15 in ref. 12, which is tailored to human subjects and assumes human-like fractionation and insensible water loss. Ape TEE values using the human-tailored equation are $11 \pm 3\%$ greater, but the pattern of TEE differences among hominoids is robust to the choice of equation (Supplementary Table 3). We prefer the more generalized mammalian equation for apes, since apes differ from humans in sweating capability and nasal morphology and are therefore likely to be more like other mammals (rather than humans) in fractionation and insensible water loss.

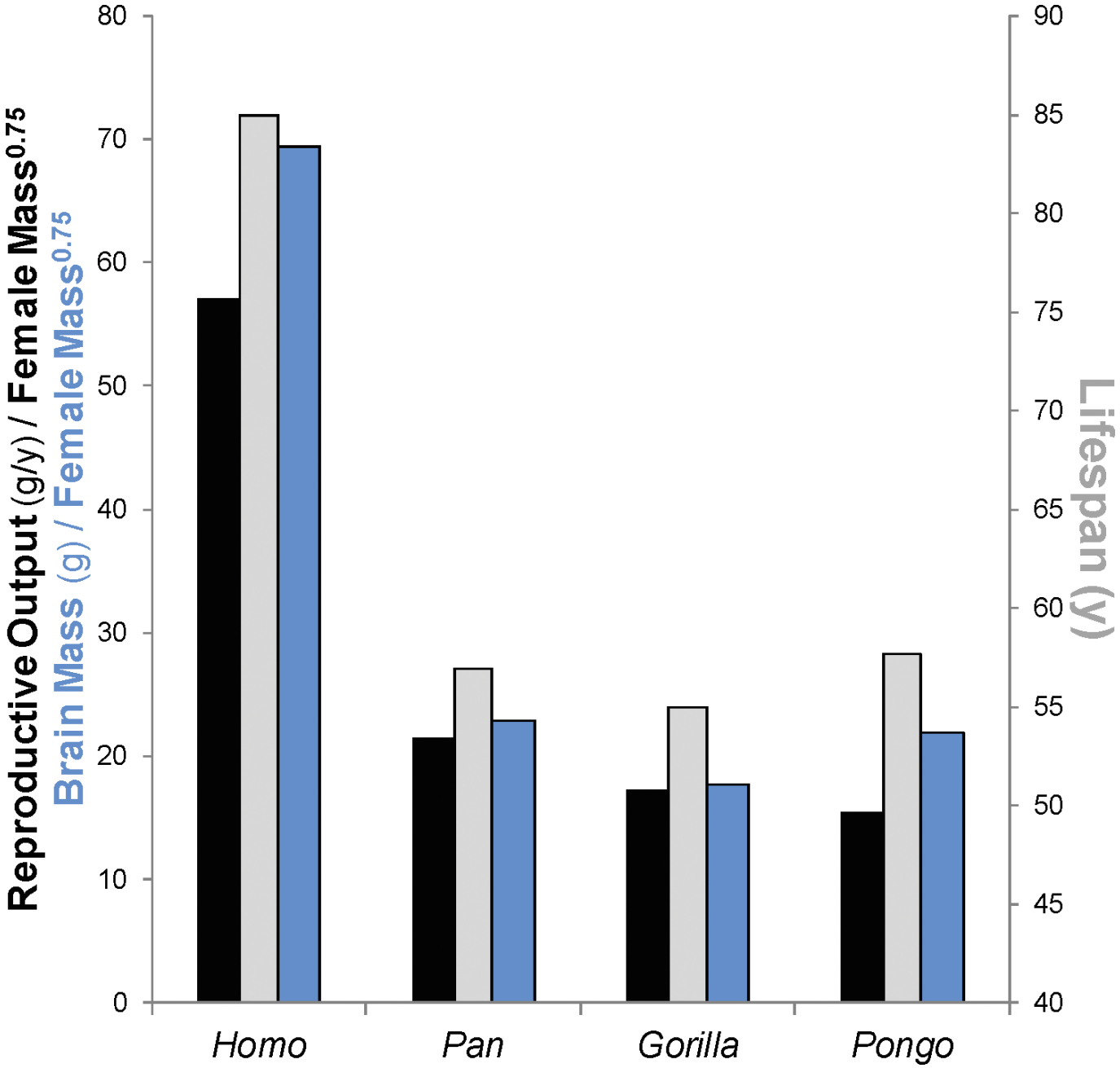
TEE analysis. Variation in TEE was examined via multiple regression[26] using ln-transformed values for TEE, FFM and fat mass to reduce heteroscedasticity of residuals. Comparisons were made at the genus level due to the close phylogenetic relatedness of bonobos and chimpanzees. Bonobos and chimpanzees did not differ ($F(32) = 0.007$, $P = 0.80$ ANCOVA) in TEE in analyses controlling for FFM.

Locomotor costs. Humans were fitted with triaxial accelerometers (Actical, Phillips Respironics) and counts per minute were logged for 7 days, corresponding with the DLW measurement period[15]. Accelerometry was not feasible for apes. Instead, physical activity was recorded using direct observation for a minimum of 20 h over the DLW measurement period. Distances walked and climbed were recorded every 15 min, using measurements of the enclosures to calibrate distance estimates. Rates of walking and climbing (metres h^{-1}) were calculated and multiplied by 12 to give the estimated distance walked and climbed during the daytime. Night-time activity was not observable in apes and was therefore excluded from analyses for all species.

BMR. Published respirometry measurements of BMR for chimpanzees[15,16] (age: 2 months to 15 years) and comparably aged humans[14] (3–18 years) were ln-transformed and analysed using multiple regression[26] controlling for sex, age, and ln-transformed body mass. Humans had higher BMRs than chimpanzees ($P < 0.001$) in all analyses (Supplementary Table 3). These data, along with the few available data for orangutan BMR[16,17], are plotted in Extended Data Fig. 2. Allometric regressions for chimpanzees and orangutans (Extended Data Fig. 2) were used to estimate BMR for the *Pan* and *Pongo* cohorts (Table 1). Published BMR prediction equations for adult humans[27] were used for the *Homo* cohort in Table 1.

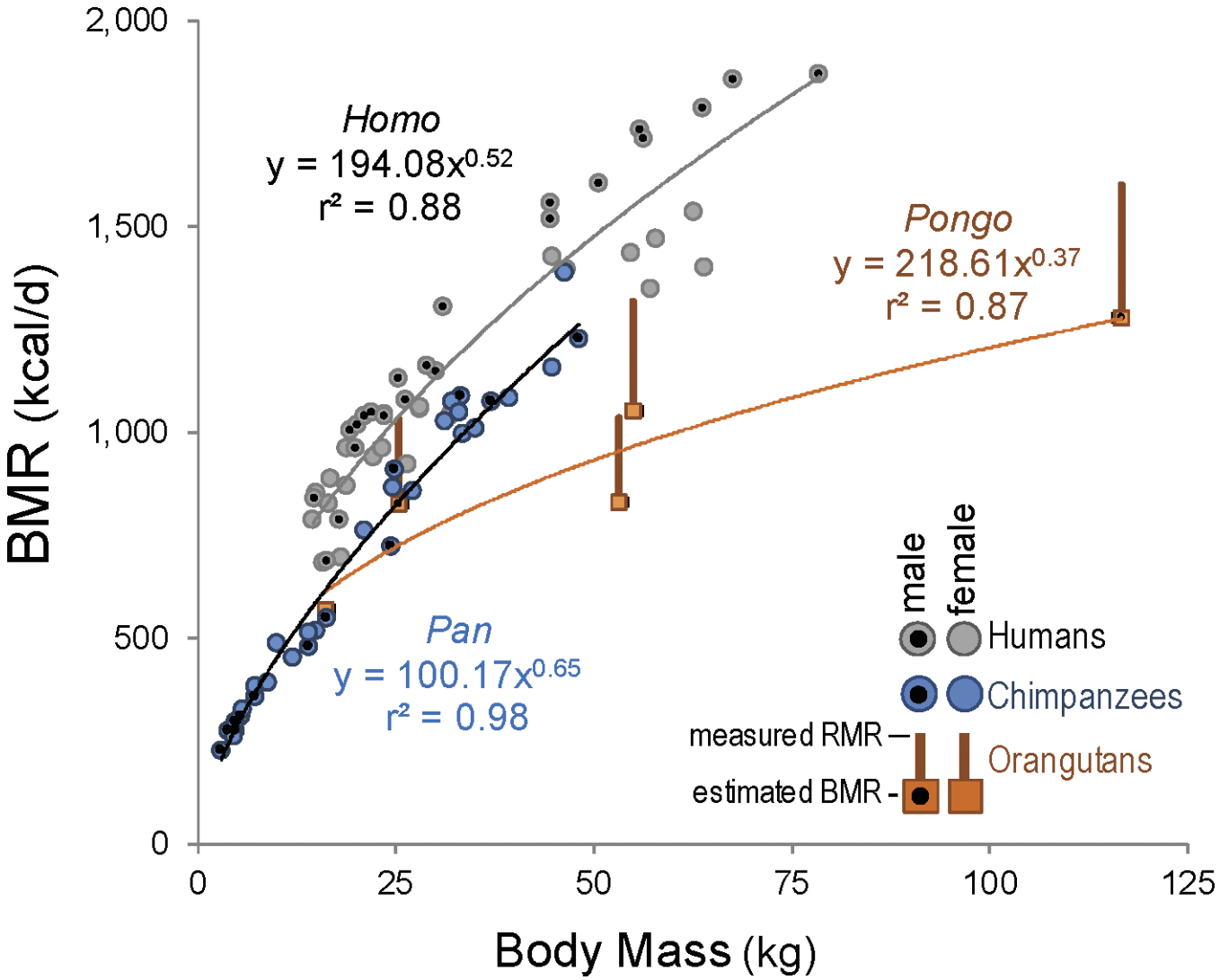
26. R Core Team. R: A language and environment for statistical computing. (R Foundation for Statistical Computing. Vienna, Austria, 2013).
27. Henry, C. J. Basal metabolic rate studies in humans: measurement and development of new equations. *Public Health Nutr.* **8,** 1133–1152 (2005).

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 1 | The human energetic paradox. Humans achieve greater reproductive output (g year^{-1}; black bar) and have larger brains (g; blue bar) relative to female metabolic mass (kg$^{0.75}$) than any of the great apes, yet also achieve longer lifespans (grey bar). Human data are from traditional hunter–gatherer and subsistence farming populations; ape data are from populations in the wild[1] (see Supplementary Table 2).

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 2 | BMRs for humans, chimpanzees and orangutans. Available BMR data for chimpanzees[15,16] are primarily from juveniles (age: 2 months to 15 years), and are shown here against comparably aged humans[14] (3–18 years). Humans have greater BMRs than chimpanzees in a general linear model of ln-transformed BMR and mass controlling for age and sex ($P < 0.001$; Supplementary Table 3). Note that in humans (but not chimpanzees), males have greater BMRs for a given body mass, reflecting their greater proportion of FFM (that is, lower body fat percentage). The available data for orangutan BMR consists of one juvenile individual[16] (no age reported, mass 16.2 kg). Resting metabolic rates (RMRs; kcal day^{-1}), measured in alert orangutans while sitting[17], are also shown. The top of the bar indicates the measured RMR for those individuals, and the bottom square indicates estimated BMR based on those measurements[17], assuming BMR = 0.8RMR. No BMR data are available for gorillas. Symbols marked with black circles are males. Allometric regressions shown for *Pan* ($y = 100.17x^{0.65}$) and *Pongo* ($y = 218.61x^{0.37}$) and were used to estimate BMR for the adult cohorts in Table 1. Human BMR in Table 1 was estimated from published, sex-specific predictive equations for adults[27].

© 2016 Macmillan Publishers Limited. All rights reserved

LETTER

doi:10.1038/nature17937

Deep-sea diversity patterns are shaped by energy availability

Skipton N. C. Woolley[1,2], Derek P. Tittensor[3,4], Piers K. Dunstan[5], Gurutzeta Guillera-Arroita[2], José J. Lahoz-Monfort[2], Brendan A. Wintle[2], Boris Worm[3] & Timothy D. O'Hara[1]

The deep ocean is the largest and least-explored ecosystem on Earth, and a uniquely energy-poor environment. The distribution, drivers and origins of deep-sea biodiversity remain unknown at global scales[1–3]. Here we analyse a database of more than 165,000 distribution records of Ophiuroidea (brittle stars), a dominant component of sea-floor fauna, and find patterns of biodiversity unlike known terrestrial or coastal marine realms. Both patterns and environmental predictors of deep-sea (2,000–6,500 m) species richness fundamentally differ from those found in coastal (0–20 m), continental shelf (20–200 m), and upper-slope (200–2,000 m) waters. Continental shelf to upper-slope richness consistently peaks in tropical Indo-west Pacific and Caribbean (0–30°) latitudes, and is well explained by variations in water temperature. In contrast, deep-sea species show maximum richness at higher latitudes (30–50°), concentrated in areas of high carbon export flux and regions close to continental margins. We reconcile this structuring of oceanic biodiversity using a species–energy framework, with kinetic energy predicting shallow-water richness, while chemical energy (export productivity) and proximity to slope habitats drive deep-sea diversity. Our findings provide a global baseline for conservation efforts across the sea floor, and demonstrate that deep-sea ecosystems show a biodiversity pattern consistent with ecological theory, despite being different from other planetary-scale habitats.

Deep-sea environments comprise approximately 66% of global sea-floor area, and hence more than half of the planet's surface[4]. The sinking of biological material to the sea floor is a critical part of the global carbon cycle and climate. Yet global patterns of sea-floor diversity remain unknown, having so far been described only on local and regional scales[4,5]. Here we assemble a unprecedented data set on the global distribution of 2,099 Ophiuroidea (brittle and basket stars) species from shallow to abyssal depths, comprising 165,044 species distribution records from 1,614 research expeditions (Extended Data Fig. 1). Ophiuroidea are an ideal model taxon to analyse global patterns of species diversity as they are a dominant component of the fauna of many deep-sea habitats[6]. These data provide a unique opportunity to uncover and compare deep-sea biodiversity patterns across three fundamentally different depth strata of the ocean, the continental shelf (20–200 m), upper continental slope (200–2,000 m) and deep-sea (2,000–6,500 m)[7,8]. Furthermore, we propose that the deep sea can be viewed as a third replicate biome (after terrestrial and shallow-water diversity) to untangle the role of fundamental processes that shape global diversity. We analysed the three bathymetric strata separately, spatially estimated and mapped total species richness across a global grid using multi-species hierarchical occupancy-detection models (MSODM), and formally tested a number of prominent hypotheses on the factors shaping deep-sea diversity patterns using spatial linear models (see Extended Data Table 1) and a species–energy framework.

Global patterns of species richness for shelf and upper-slope species are congruent with those of coastal marine species[9]. Both communities show diversity peaks in the tropical Indo-west Pacific and the western Atlantic Oceans (Fig. 1a, b). However, in contrast to previous work, we find relatively high regional species richness around southern Australia and New Zealand (Fig. 1b)[9]. Species richness is generally suppressed on the western side of tropical America and Africa, and the northern Indian Ocean (Fig. 1a, b and Extended Data Fig. 2a, b). Deep-sea species richness shows a markedly different pattern, with peaks occurring predominantly at mid-to-high latitudes (Fig. 1c and Extended Data Fig. 2c), particularly across the boreal Atlantic Ocean, around Japan, New Zealand, western North and South America and western Africa.

When global ophiuroid richness is examined by latitude and depth (Fig. 2), it peaks in the tropics at continental shelf (20–200 m) and upper-slope depths (200–1,200 m). A strong latitudinal biodiversity gradient exists at these depths with reduced richness at mid-to-high latitudes (>45° S and >55° N). These results are congruent with prior studies of shallow-water[9] and terrestrial[10] global diversity gradients that suggest a unimodal diversity peak at low latitudes. However, at lower-slope to intermediate abyssal depths (2,000–6,500 m), bimodal maxima occur at temperate latitudes (30–40° S and 40–50° N), with distinct minima near the poles and at northern subtropical latitudes (15–30° N). Thus the typical latitudinal gradient of species richness observed near the planetary surface does not hold in the deep sea.

We encode a priori hypotheses on processes expected to structure biodiversity[11] by encapsulating them as potential drivers in a spatially explicit statistical model (see list of hypotheses in Extended Data Table 1). Geographical variation in energy availability (the species–energy hypothesis) is a factor thought to shape terrestrial and marine global biodiversity[11,12], through radiation (light), thermal (kinetic) or chemical (potential) energy. Unlike other realms, the first of these can be excluded from the aphotic deep-sea environment. Thermal energy may affect diversity through several mechanisms, including physiological tolerances, speciation/extinction rates, and availability of metabolic niches[13]. Chemical energy in the form of reduced organic compounds is hypothesized to promote species diversity[13]; in the deep-sea this would be reflected by food resource availability manifested as particulate organic carbon (POC) flux. Non-energetic factors tested included oxygen stress, reflected on the upper slope by oxygen minimization zones (OMZs)[14]; the environmental stress hypothesis proposes that species richness has a negative relationship with environmental stress[15]. Finally, long-term connectivity between shallower shelf and upper-slope species to deep-sea communities is expected to affect species richness[16] via the regulation of deep-sea populations through extinction and radiation of species from connected regions[17]. Testing these hypotheses against patterns of deep-sea diversity helps

[1]Museum Victoria, GPO Box 666, Melbourne, Victoria 3001, Australia. [2]Quantitative and Applied Ecology Group, School of Biological Sciences, BioSciences Building 2, The University of Melbourne, Victoria 3010, Australia. [3]Department of Biology, Dalhousie University, 1355 Oxford Street, Halifax B3H 4J1, Canada. [4]United Nations Environment Programme World Conservation Monitoring Centre, 219 Huntingdon Road, Cambridge CB3 0DL, UK. [5]CSIRO, Wealth from Oceans Flagship, Hobart, Tasmania 7000, Australia.

© 2016 Macmillan Publishers Limited. All rights reserved

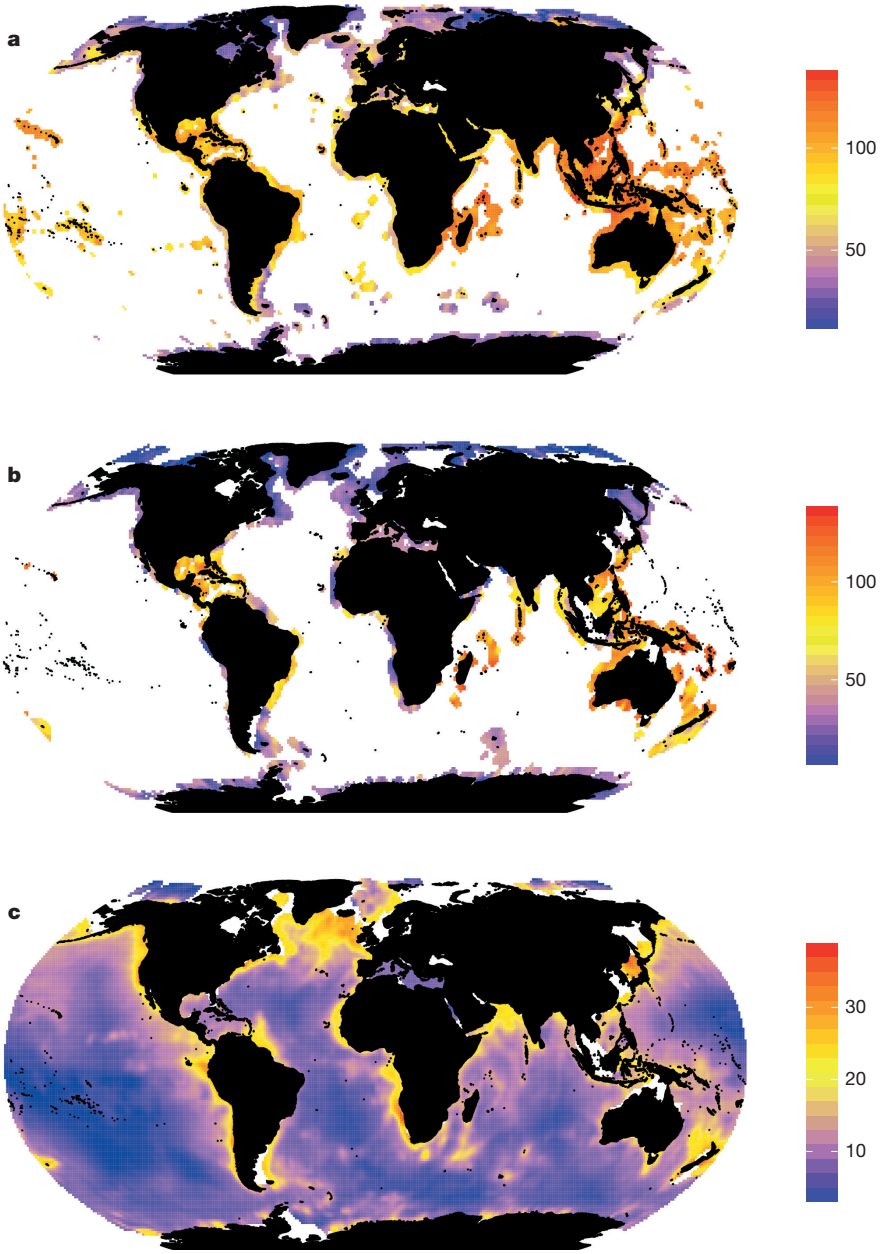


Figure 1 | Global patterns of ophiuroid species richness. a–**c**, Multi-species occupancy detection models (MSODM) of summed occupancy probabilities at 500 km equal area resolution for shelf diversity (20−200 m; **a**), slope diversity (200−2,000 m; **b**) and deep-water diversity (2,000−6,500 m; **c**). Land maps used in these figures are from the public domain Natural Earth and freely available for personal, educational, and commercial purposes. See http://www.naturalearthdata.com/about/terms-of-use/ for more details.

disentangle the environmental, ecological and historical forces shaping global diversity.

Our statistical models revealed that the species–energy hypothesis is broadly supported at all depths, albeit through different forms of energy (Table 1). A significant relationship ($P < 0.01$) between richness and sea-floor water temperature emerges at shelf and upper-slope depths, correlating with kinetic (specifically thermal) energy input from the sun. Strong thermal gradients are present in shelf and slope (but not deep-sea) regions, promoting greater species richness[13,18]. However, there is a significant negative correlation between the diversity of shelf and slope environments and chemical energy, measured as POC flux to the sea floor (Table 1; POC, $P < 0.05$), probably because tropical shallow-water systems tend to be nutrient-poor. Conversely, deep-sea richness is not correlated with temperature but is significantly positively correlated with chemical energy export (Table 1; POC, $P < 0.01$) and regions of high seasonal surface productivity (SVI; $P < 0.01$). POC export is probably a key source of energy that maintains deep-sea species in regions of constant and low thermal energy[19–21]. The diversity of shelf communities within the model is also suppressed in OMZs (Table 1; oxygen stress[2], $P < 0.05$). These zones are dysoxic, with less than 2 ml O_2 per litre of seawater, and are known to have substantially reduced faunal diversity and biomass[14]. Thus environmental stress appears to play an additional role in influencing global patterns of ophiuroid richness. For deep-sea environments, connectivity is also a significant predictor, with a decline in species diversity with distance from continental margins (Table 1; DC, $P < 0.05$). This finding implies that the continental margins are a long-term source of abyssal diversity. This is consistent with the radiation hypothesis[16] that predicts deep-water diversity is maintained by immigration from bathyal sources[17].

© 2016 Macmillan Publishers Limited. All rights reserved

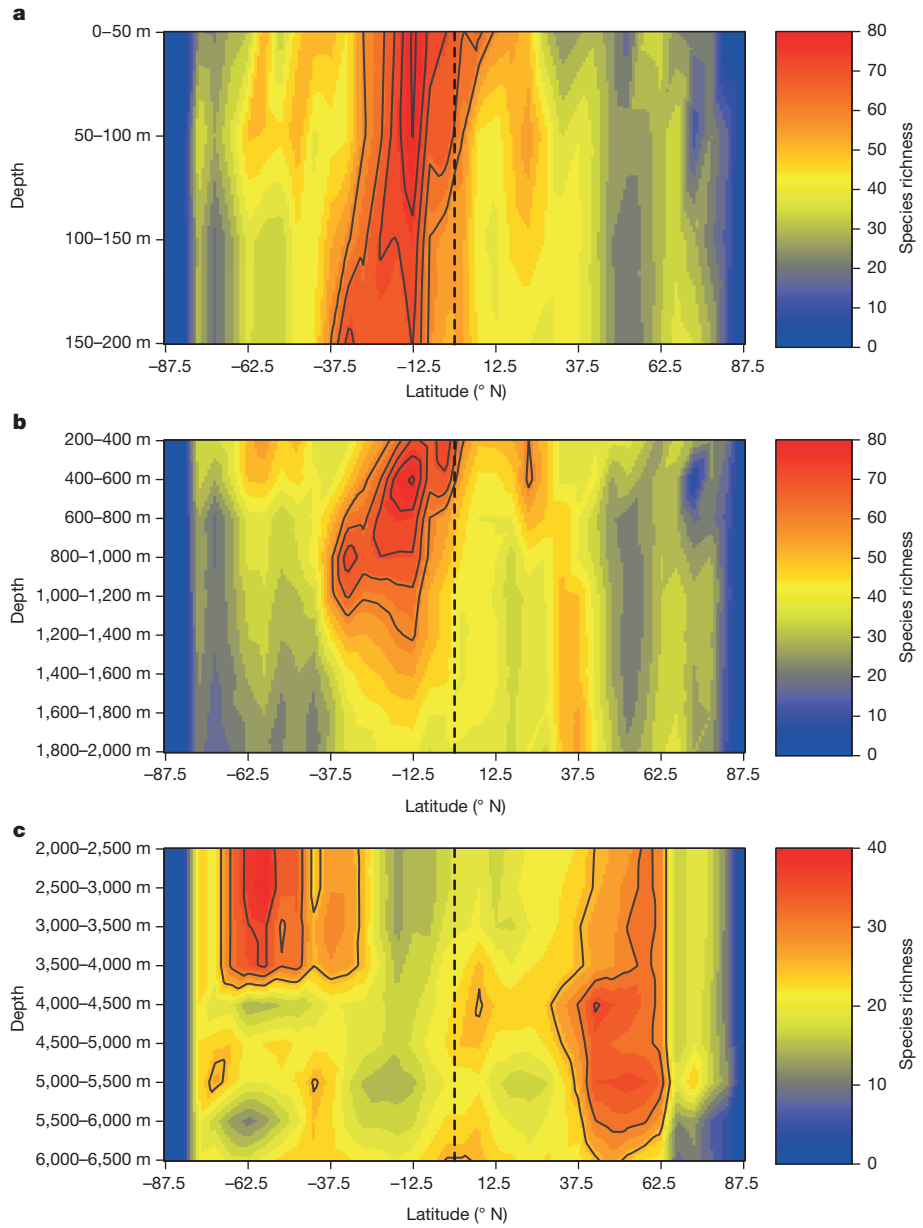


Figure 2 | Estimated mean ophiuroid species richness plot as a function of depth and latitude. a–**c**, Species richness predicted from MSODMs at depth intervals from surface to lower-abyss depths for binned equal-area latitudinal regions across the global extent of longitude. Mean species richness estimated from MSODMs for shelf diversity (20–200 m; **a**), slope diversity (200–2,000 m; **b**) and deep-water diversity (2,000–6,500 m; **c**). The vertical dashed line represents the equator. The grey contour lines represent the top 20% of species richness for each bathome. Depths are binned at 50 m intervals for shelf regions (0–50, 50–100, 100–150 and 150–200 m), at 200-m intervals for slope (200–2,000 m), and 500-m intervals for deep-water (2,000–6,500 m).

Observed patterns of deep-water diversity may shed some light on the drivers of large-scale gradients of diversity in other environments. Mannion *et al.*[22] suggested that two different classes of latitudinal diversity gradients occurred over the history of the Phanerozoic eon (542 million years ago). A tropical maxima and polar minima existed during relatively cool 'icehouse' conditions[23] where there is a strong sea surface temperature divergence between equator and pole. Conversely, a flat diversity gradient or temperate peak occurred during warmer 'greenhouse' conditions, when there was less of a temperature gradient, indicating that thermal energy was probably a key driver of geographic variation in richness. Here we observe no tropical peak in diversity of deep-water assemblages (2,000–6,500 m), suggesting that it is uniform temperatures, rather than hot 'greenhouse' conditions, that supress a low-latitude diversity peak. Our results imply that energy availability determines the latitudinal diversity gradient, but in the deep sea, unlike the rest of the ocean, this derives from chemical rather than thermal sources.

Currently, we know little about the evolutionary processes that at least partially gave rise to bathyal and abyssal species diversity patterns. Multiple hypotheses postulate *in situ* diversification, or immigration from shallower depths[16,24]. It is unclear how the abyssal and lower bathyal fauna re-establish after extinction events such as anoxia[25]. Our modelled estimates of species richness depict higher diversity on the upper slope; these results lend support to the upper slope being the source of deep-sea diversification. Observed patterns of species richness are highest near continents (Fig. 1), showing the relationship between deep-water diversity and connectivity to continental margins (Table 1). Thus our data and analyses lend support to the theory that the deep-sea fauna at least partially originates from range expansion of upper-bathyal species into the deep-sea. However, these suggested peaks of diversity

© 2016 Macmillan Publishers Limited. All rights reserved

Table 1 | SLM for species richness of three bathomes

Bathome	20–200 m (shelf)	200–2,000 m (upper slope)	2,000–6,500 m (lower slope and abyss)
Species richness	126	110	31
Annual mean temperature (AMO)	11.49***	3.61***	
Annual mean temperature (AMO)^2		−2.71**	
Annual mean oxygen (AMO)	−2.17*		
Annual mean oxygen (AMO)^2			
Seasonal variation in NPP (SVI)	3.54**		1.61**
Seasonal variation in NPP (SVI)^2	−2.48*		
Particulate organic carbon flux (POC)	−4.43*	−3.06*	3.09**
Particulate organic carbon flux (POC)^2	2.13*		−2.46*
Distance to continental margin (DC)	n/a	n/a	0.45*
Distance to continental margin (DC)^2	n/a	n/a	
Oxygen stress (OMZ)			
Oxygen stress (OMZ)^2	1.71*		
Pseudo-R^2	0.35	0.37	0.21

Spatial linear model (SLM) results for the species richness of the three bathomes 20–200 m, 200–2,000 m and 2,000–6,500 m are shown. Maximum species richness for each bathome is highest individual cell value. Model results are from the best SLM as determined by Akaike information criterion value. Model results are *z* values; $*P < 0.05$, $**P < 0.01$, $***P < 0.0001$. Distance to continental margin is only applicable for deep-water (LSA; lower-slope and abyss). n/a, not applicable.

are also regions of high export productivity, a strong energetic predictor of richness. To delineate better the processes shaping evolutionary origins of deep-sea fauna, comprehensive phylogenies are required.

In conclusion, our findings reveal a unique global pattern of deep-sea benthic biodiversity that is unlike any other environment. We are able to reconcile the vertical structuring of marine biodiversity through a species–energy framework, a fundamental theory of the origination of biodiversity. Our findings also support the radiation hypothesis, suggesting deep-water richness is maintained by immigration from shallower regions. These results have important implications for identifying potential protected areas on the high seas, both within and outside national jurisdictions. Tropical areas are typically highly diverse in shallow waters and on land, and thus often the focus for conservation efforts. In contrast, areas of higher export productivity and connectivity to shallower communities will need to be considered for conservation and management action in the deep-sea. Our results provide a much-needed empirical and spatial baseline for global conservation planning in the deep ocean, which is urgently motivated by the accelerating pressures from deep-sea fishing, mining and other cumulative impacts on this vulnerable region[26].

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 24 January 2015; accepted 21 March 2016.
Published online 11 May 2016.

1. Brandt, A. *et al.* First insights into the biodiversity and biogeography of the Southern Ocean deep sea. *Nature* **447,** 307–311 (2007).
2. Rex, M. A. *et al.* Global-scale latitudinal patterns of species diversity in the deep-sea benthos. *Nature* **365,** 636–639 (1993).
3. Poore, G. C. B. & Wilson, G. Marine species richness - Reply. *Nature* **361,** 598 (1993).
4. Ramirez-Llodra, E. *et al.* Deep, diverse and definitely different: unique attributes of the world's largest ecosystem. *Biogeosciences* **7,** 2851–2899 (2010).
5. Rex, M. A. & Etter, R. J. *Deep-Sea Biodiversity; Patterns and Scale* (Harvard Univ. Press, 2010).
6. Stöhr, S., O'Hara, T. D. & Thuy, B. Global diversity of brittle stars (Echinodermata: Ophiuroidea). *PLoS One* **7,** e31940 (2012).
7. Watling, L., Guinotte, J., Clark, M. R. & Smith, C. R. A proposed biogeography of the deep ocean floor. *Prog. Oceanogr.* **111,** 91–112 (2012).
8. O'Hara, T. D., Rowden, A. A. & Bax, N. J. A southern hemisphere bathyal fauna is distributed in latitudinal bands. *Curr. Biol.* **21,** 226–230 (2011).
9. Tittensor, D. P. *et al.* Global patterns and predictors of marine biodiversity across taxa. *Nature* **466,** 1098–1101 (2010).
10. Kreft, H. & Jetz, W. Global patterns and determinants of vascular plant diversity. *Proc. Natl Acad. Sci. USA* **104,** 5925–5930 (2007).
11. Currie, D. J. *et al.* Predictions and tests of climate-based hypotheses of broad-scale variation in taxonomic richness. *Ecol. Lett.* **7,** 1121–1134 (2004).
12. Wright, D. H. Species-energy theory: an extension of species-area theory. *Oikos* **41,** 496–506 (1983).
13. Clarke, A. & Gaston, K. J. Climate, energy and diversity. *Proc. R. Soc. Lond. B* **273,** 2257–2266 (2006).
14. Levin, L. A. Oxygen minimum zone benthos: adaptation and community response to hypoxia. *Oceanogr. Mar. Biol.* **41,** 1–45 (2003).
15. Fraser, R. H. & Currie, D. J. The species richness-energy hypothesis in a system where historical factors are thought to prevail: coral reefs. *Am. Nat.* **148,** 138–159 (1996).
16. Holt, R. D. Population dynamics in two-patch environments: Some anomalous consequences of an optimal habitat distribution. *Theor. Popul. Biol.* **28,** 181–208 (1985).
17. Rex, M. A. *et al.* A source-sink hypothesis for abyssal biodiversity. *Am. Nat.* **165,** 163–178 (2005).
18. Rohde, K. Latitudinal gradients in species diversity: the search for the primary cause. *Oikos* **65,** 514–527 (1992).
19. Tittensor, D. P., Rex, M. A., Stuart, C. T., McClain, C. R. & Smith, C. R. Species-energy relationships in deep-sea molluscs. *Biol. Lett.* **7,** 718–722 (2011).
20. Yasuhara, M. & Danovaro, R. Temperature impacts on deep-sea biodiversity. *Biol. Rev.* **91,** 275–287 (2014).
21. Lambshead, P. J. D., Tietjen, J., Ferrero, T. & Jensen, P. Latitudinal diversity gradients in the deep sea with special reference to North Atlantic nematodes. *Mar. Ecol. Prog. Ser.* **194,** 159–167 (2000).
22. Mannion, P. D., Upchurch, P., Benson, R. B. J. & Goswami, A. The latitudinal biodiversity gradient through deep time. *Trends Ecol. Evol.* **29,** 42–50 (2014).
23. Smith, A. G. & Pickering, K. T. Oceanic gateways as a critical factor to initiate icehouse Earth. *J. Geol. Soc. Lond.* **160,** 337–340 (2003).
24. Rex, M. A., Crame, J. A., Stuart, C. T. & Clarke, A. Large-scale biogeographic patterns in marine mollusks: a confluence of history and productivity? *Ecology* **86,** 2288–2297 (2005).
25. Rogers, A. D. The role of the oceanic oxygen minima in generating biodiversity in the deep sea. *Deep Sea Res. Part II Top. Stud. Oceanogr.* **47,** 119–148 (2000).
26. Ramirez-Llodra, E. *et al.* Man and the last great wilderness: human impact on the deep sea. *PLoS One* **6,** e22588 (2011).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank all collectors and taxonomists involved for providing biodiversity data. This work is an output of the project 'National maps of biodiversity and connectivity' of the Marine Biodiversity Research Hub and Environmental Decisions Hub, funded through the Commonwealth National Environmental Research Program (NERP) and administered through the Australian Government's Department of Environment. This work is also a product of the International Network for Scientific Investigations of Deep-Sea Ecosystems (INDEEP) working group on biogeography. We also thank the Centre of Excellence for Environmental Decisions (CEED) for travel funding that enabled collaboration between the University of Melbourne, Museum Victoria and Dalhousie University.

Author Contributions S.N.C.W., T.O., D.T. and B.W. conceived the study. T.O. collected, refined and managed the biological dataset. S.N.C.W., T.O., D.T., B.A.W., G.G.A. and J.J.L.M. performed analyses. All authors contributed to writing the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.N.C.W. (swoolley@museum.vic.gov.au).

© 2016 Macmillan Publishers Limited. All rights reserved

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Data. *Biological data.* Global brittle-star occurrence data (84° N to 78° S latitude and 180° W to 180° E longitude) was derived from 1,614 research expeditions, covering a 130 year timespan, starting with iconic nineteenth century voyages such as the *Challenger* expedition[27]. Brittle-star species occurrence records were collected from three major bathomes: shelf (SH; 20−200 m), slope (SL, 200−2,000 m) and deep-water (lower slope and abyssal plane; LSA, 2,000−6,500 m). These depth strata were selected to reflect existing biogeographical bounds of bathyal ophiuroids[8]. Ophiuroid occurrences at hadal depths (>6,500 m) were removed as the data were very sparse and would probably result in fragile inference of patterns at these depths. Ophiuroidea identifications were verified by taxonomic experts to species level (including author T.O.). Specimens were collected using *ad hoc*, semi-quantitative and quantitative methods, including trawls, dredges, epibenthic sleds, grabs and corers. The highest density of *ad hoc* samples (for example, collections by hand) corresponded to coastal occurrence records (0−20 m) and were subsequently removed from analyses to minimise potential collection bias[28]. The spatial extent and proportion of survey effort per 500-km cell were plotted to visualize spatial bias in collection effort (Extended Data Fig. 1).

Environmental data. Environmental and physical predictors were used to test hypotheses that seek to explain patterns of deep-sea species richness (see Extended Data Table 1 for a summary of hypotheses name, meaning, relevance, origin and related predictors). Environmental predictors were tri-linearly interpolated to the sea floor using global ETOPO1 ice-surface GIS bathymetric data set[29]. Annual mean sea-floor temperature (°C) (AMT), annual standard deviation of sea-floor temperature (°C) (ASDT) and annual mean oxygen (ml l^{-1}) (AMO) were derived from the CARS 2009 data set[30,31]. The CARS climatology physical oceanography data (1950–2009) were interpolated across the globe for 79 depth layers at a resolution of 0.5° latitude/longitude. We also calculated the proportion of AMO grid cells that had <2 ml per litre O_2, a critical physiological limit for numerous marine species[32] and typically the threshold for oxygen minimization zones (OMZs)[14]. Mean annual net primary productivity (g C m^{-2} $year^{-1}$, NPP) and seasonal variation of net primary productivity (g C m^{-2} $year^{-1}$; SVI) were generated from a vertically generalized production model (VGPM)[33]. NPP and SVI are a function of satellite-derived chlorophyll (SeaWiFS). NPP and SVI were calculated across the years 2003 to 2010 (see http://www.science.oregonstate.edu/ocean.productivity/). Particulate organic carbon flux to the sea floor (POC flux; g C m^{-2} $year^{-1}$) was estimated using NPP and SVI data and a productivity export model[34]. Distance from continental margins (DC) for deep-sea habitats was estimated based on the IFREMER Continental margins shape file[35]. We created a spatial layer that calculates distance of sea-floor habitat to nearest point on the 2,000 m contour around continental margins and islands.

For the statistical analyses, environmental predictors were averaged to cell size across the three bathomes (20–200 m, 200–2,000 m and 2,000–6,500 m). Strongly correlated variables (>0.7) were removed from analyses to avoid issues with co-linearity of model coefficients (Extended Data Table 2). AMO was removed from the shelf analysis, due to its correlation with AMT. We removed NPP from analyses due to its correlation with POC flux. We selected POC flux over NPP as we were interested in the amount of carbon flux at the sea floor, rather than the surface. All independent variables used in statistical analyses were centred and normalized (mean = 0, variance = 1). All analyses were performed at spatial scales of 500 km equal area grid cells.

Statistical analysis. We were interested in describing patterns of species richness and the processes that shape observed patterns in the deep-sea benthos. Many authors have approached these analyses using either bottom-up[36,37] or top-down methods[10]. The respective merits of both approaches are still debated in the ecological literature[38]. We see merits in both approaches, using them for different purposes.

For a top-down approach we linked estimates of species richness derived from the estimated asymptotes of species accumulation curves to environmental and physical data using spatial regression models (SLMs). This assumes that the environment is likely to impose top-down limits of species richness, independently of species identities. Critically, unlike the species distribution modelling (see below), we estimated the number of species in a region (cell) independently of environment predictors, thus enabling us to assess potential determinants of richness in our modelling framework without circular reasoning. We therefore use this approach to test hypotheses of processes that shape global deep-sea species richness.

Our second approach used the summation of species distribution models to assess species richness. Here we used an extension of classic species distribution models that incorporates detection probabilities when assessing the distribution of modelled species. Details of our two approaches are discussed in the following section.

Spatial linear models (SLMs). *Estimation of richness via species accumulation curves.* The 'coverage-based rarefaction' method[39] was used to estimate species richness on a cell-by-cell basis. This method estimates species richness based on a measure of sample completeness[39]. The aim is to estimate the 'sample deficit', which represents the fraction of the community that remains undiscovered[40]. This is a novel alternative to species accumulation curves based on the extrapolation of individuals or samples[41], and one which attempts to scale the richness of each cell to an equivalent level of sampling coverage for all cells. For our maps of diversity, we used a 75% coverage-based estimate of the number of species per cell as a conservative balance between extrapolation and completeness of sample coverage. For each cell we ran the estimator with 1,000 bootstraps, and took the mean as our point estimate of species richness for each cell. To assess the performance of cell-by-cell estimation of the number of species, we plotted all estimated species accumulation curves and their respective bootstrap bounds for each cell and visually assessed the curvilinear nature of each extrapolation. Species accumulation curves that (1) did not show asymptotic behaviour or (2) had extreme confidence bounds based on bootstrapping, were removed from further analysis. Species richness interpolations and extrapolations were calculated using the R package Vegan[42] and code adapted from the package iNEXT[43].

Modelling of estimated richness as a function of environmental predictors. Estimated species richness was used as a response variable in models that tested hypotheses about its relationship to environmental predictors. We used SLMs that explicitly account for spatial autocorrelation (Extended Data Table 3), specifically simultaneous autoregressive models (SARs)[44]. Neighbourhood size was selected using an error-SAR process, based on the minimum AIC for spatial null models (model containing the intercept and the spatial autocorrelation term). Neighbourhood sizes between 1,000 km and 10,000 km were tested at 100-km intervals. Neighbourhood size was determined independently for each depth strata, as it was expected that different bathomes would display differing extents of spatial autocorrelation due to different ecological and evolutionary process driving the spatial patterns of species richness. We used an all-model selection method to find the AIC-best model. We analysed the models and the relative importance of predictors through *z*-tests (SLMs). We used pseudo-R^2 to assess model fit. We fitted linear, and second-order polynomial functions for each predictor variable given that a number of studies have emphasized the importance of unimodal relationships with temperature[45] and POC flux[9] (Extended Data Fig. 3). Models were fitted using the errorslm function in spdep[46] package in R.

Multi-species occupancy-detection models (MSODMs). The second approach for analysis involved MSODMs[47–49], a relatively novel but promising community-modelling framework that allows flexible consideration of species distributions and their detectability. This modelling framework is grounded in the view that species richness and other attributes of community structure are best described using models of individual species occurrence that explicitly account for imperfect detection during sample collection[50,51]. This framework thus explicitly deals with potential biases in sampling effort, as those expected in our deep-sea species data.

MSODMs provide a hierarchical and explicit description of the state (species occurrences) and observation (species detection) processes. At the heart of the approach is the estimation of the incompletely observed site-by-species occurrence matrix, from which different summaries of community structure can be derived. The presence or absence of species *i* at a site *j* is described as the outcome of a Bernoulli trial

$$Z_{ij} \sim \text{Bernoulli}(\psi_{ij})$$

where ψ_{ij} is the probability that species *i* is present at site *j*, and the latent variable Z_{ij} represents whether the species is present or not at the site (Z_{ij} takes value 0 or 1).

The observation model describes the observed data as the outcome of a series of independent Bernoulli trials with probability p_{ijk} at sites where the species is present ($Z_{ij} = 1$) and 0 elsewhere, that is,

$$Y_{ijk}|Z_{ij} \sim \text{Bernoulli}(Z_{ij}\, p_{ijk})$$

where Y_{ijk} are the observed data (detection/non-detection of species *i* at site *j* during survey *k*), and p_{ijk} are the corresponding species detection probabilities (the probability of detecting species *i* at site *j* during survey visit *k*). The model assumes that the occupancy status of cells (Z_{ij}) do not change during the survey period, which is a reasonable assumption at the geographical scale of our analysis and time frame of the data collection. Occupancy and detection probabilities can

© 2016 Macmillan Publishers Limited. All rights reserved

then be modelled as a function of relevant environmental predictors following the generalized linear modelling framework, for example,

$$\text{logit}(\psi_{ij}) = \beta_{0i} + \beta_{1i} \times \text{covariate}_{1j} + \ldots + \beta_{ni} \times \text{covariate}_{nj}$$

where n is here the number of predictors in the occupancy component of the model (including quadratic terms, interactions, and so on). In our model, occupancy probability was described as a function of the 12 to 14 covariates (depending on the depth strata) using linear and quadratic terms (Extended Data Figs 4 and 5). We ran a single model with all covariates and considered covariate contribution, rather than using model selection. Detection probability was described as a function of the collection method (for example, dredge or grab) used in each collection event (that is, survey visit k at site j)

$$\text{logit}(p_{ijk}) = \alpha_{oi} + \alpha_{1i} \times \text{gear}_{jk}$$

In the MSODM framework, individual species models are linked through random effects in a hierarchical fashion, this way exploiting similarities in environmental responses to borrow information across species. This is achieved by describing the parameters from species-specific models as realizations from a common distribution, whose parameters (or 'hyperparameters') are estimated. For our analysis, parameters were described using independent normal distributions as follows

$$\beta_{xi} \sim N(\mu_{\beta_x}, \sigma^2_{\beta_x}) \text{ and } \alpha_{xi} \sim N(\mu_{\alpha_x}, \sigma^2_{\alpha_x})$$

Once a MSODM is fit, species richness and other metrics of community structure can be derived based on the parameter estimates obtained. In particular, species richness is simply obtained by summing the estimated occupancy probabilities across species. The estimated species richness at site $j(\hat{N}_j)$ is thus calculated as

$$\hat{N}_j = \sum_{i=1}^{I} \hat{\psi}_{ij}$$

that is, the species richness estimate at site j is equal to the expected number of species at the site.

The MSODM framework allows inference about the number of species that were completely missed during sampling[47,48]. In our analysis however, estimation was restricted to species with a minimum of ten occurrences because of computational limitations on estimating rare and undetected species in large species by sites matrices. We fitted separate MSODMs to the three bathomes, shelf (320 species), upper slope (440) and deep-sea (58) (Fig. 1a–c).

Our MSODM model makes two key assumptions: cell sites are visited multiple times over a period of population closure during which the occupancy status (Z_{ik}) of a site does not change. This assumption is likely to be violated at small spatial scales, however as we are estimating occupancy at large scales, it is possible to assume that occupancy in the 500-km cells should not change over the time scale of these data collections. The second assumption is that species identification is constant across the data set with no false-positive identifications in the data. The data used in this analysis was verified by a taxonomic expert (T.O.) from museum-based records or the taxonomic literature to ensure consistent identification across the whole data set. The spatial MSODMs for shelf, slope and deep-water assemblages are presented in Fig. 1a–c. We can also assess the uncertainty in spatial predictions of occupancy probabilities across all species in the MSODMs, we do this by presenting the mean variance in MSODM predictions for shelf, slope and deep-water diversity. The variance predicted occupancy probability is calculated spatially for each species and the mean variance across all species per cell (Extended Data Fig. 6a, b).

Models were fitted using JAGS[52,53], a program for Bayesian inference using Markov chain Monte Carlo (MCMC). JAGS was controlled via an R script using package R2jags[54]. Three chains were run with different initial values, a burn-in of 2,000 iterations and a minimum of 20,000 iterations with a thinning by 50. Model convergence was assessed using the $\hat{R}$ ('R-hat') statistic[55]. We present parameter posteriors distributions for covariate estimates, which represent the distribution of all species response to each covariate. We also present the 10th–90th percentile of species partial response to covariates as a function of occupancy (ψ) for each bathome (Extended Data Fig. 4), we also report the mean posterior distributions of parameter estimates for all species (Extended Data Fig. 5).

We also compared deviances of null (intercept only) and full covariate models for each bathome. Table of deviances, DIC and pD (an estimate of deviance relative to variance) are presented in Extended Data Table 4. Bespoke C++ code, written using C++ and Armadillo C++ linear libraries[56], which was integrated into the R environment using Rcpp[57] and RcppArmadillo[58], was used to predict the occupancy of species individually from fine-scale environmental data. The code uses a 500 km moving window to estimate fine scale probabilities based on the original 500 km resolution of the original MSODMs. This essentially smooths predictions to be representative of the original cell size estimates.

Figure 2 was derived from MSODMs, by predicting the probability of occupancy for each species at a series of depth bounds. For the shelf we predicted the probability of our 320 species at 50-m depth intervals (0–50, 50–100,100–150 and 150–200 m). For slope species we broke up the environment into 200-m depth intervals (200–2,000 m). While for abyss and lower slope we broke up the depth bands into 500-m intervals (2,000–6,500 m). The estimated species richness at site $j(\hat{N}_j)$ is thus calculated for each 500-km cell at each depth layer. We then took the mean of $\hat{N}_j$ for each latitudinal band across the global prediction.

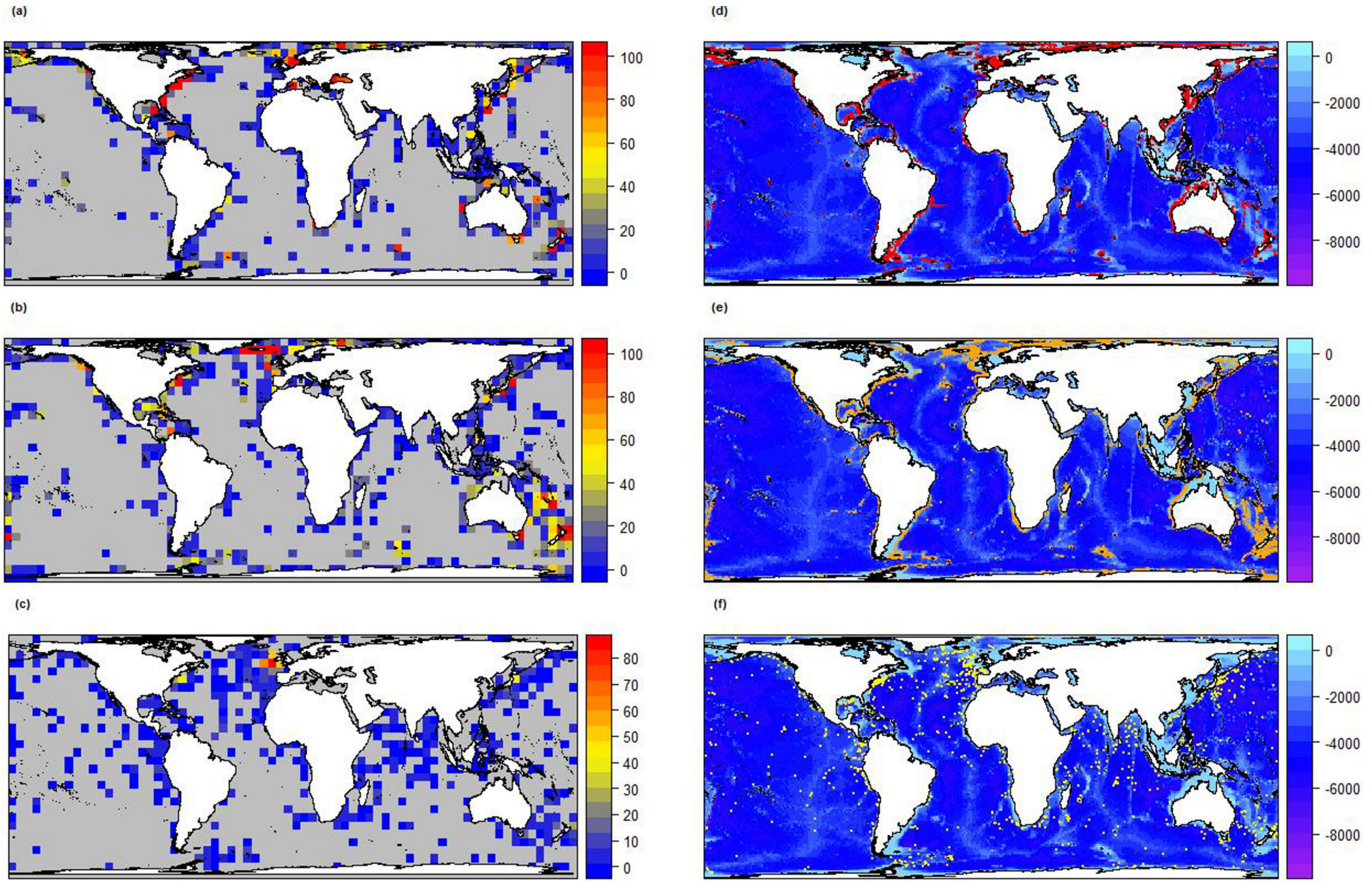
All analyses were undertaken in the R statistical language version 3.0 (ref. 59). Details about the packages and functions used are given under each section (we have made our JAGS model code publicly available on GitHub: https://github.com/skiptoniam/msodm_code). Spatial predictions were plotted using ArcMap v10 (ref. 60) and R spatial packages (Rgdal[61], raster[62], Maptools[63] and dismo[64]).

27. Murray, J. & Hjort, J. *The Depths of the Ocean: A General Account of the Modern Science of Oceanography Based Largely on the Scientific Researchs of the Norwegian Steamer Michael Sars in the North Atlantic* (Macmillan, 1912).
28. Graham, C. H., Ferrier, S., Huettman, F., Moritz, C. & Peterson, A. T. New developments in museum-based informatics and applications in biodiversity analysis. *Trends Ecol. Evol.* **19,** 497–503 (2004).
29. Amante, C. & Eakins, B. W. ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis. *NOAA Technical Memorandum NESDIS NGDC-24* (2009).
30. Ridgway, K. R., Dunn, J. R. & Wilkin, J. L. Ocean interpolation by four-dimensional weighted least squares - application to the waters around Australasia. *J. Atmos. Ocean. Technol.* **19,** 1357–1375 (2002).
31. Dunn, J. R. & Ridgway, K. R. Mapping ocean properties in regions of complex topography. *Deep Sea Res. Part I Oceanogr. Res. Pap.* **49,** 591–604 (2002).
32. Vaquer-Sunyer, R. & Duarte, C. M. Thresholds of hypoxia for marine biodiversity. *Proc. Natl Acad. Sci. USA* **105,** 15452–15457 (2008).
33. Behrenfeld, M. J. & Falkowski, P. G. Photosynthetic rates derived from satellite-based chlorophyll concentration. *Limnol. Oceanogr.* **42,** 1–20 (1997).
34. Lutz, M. J., Caldeira, K., Dunbar, R. B. & Behrenfeld, M. J. Seasonal rhythms of net primary production and particulate organic carbon flux to depth describe the efficiency of biological pump in the global ocean. *J. Geophys. Res.* **112,** C10 (2007).
35. IFREMER. (Vion, A.; Menot, L.). Continental margins between 140m and 3500m depth. http://www.marineregions.org/ (2009).
36. Guisan, A. *et al.* Making better biogeographical predictions of species' distributions. *J. Appl. Ecol.* **43,** 386–392 (2006).
37. Austin, M. P. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecol. Modell.* **157,** 101–118 (2002).
38. Boucher-Lalonde, V., Kerr, J. T. & Currie, D. J. Does climate limit species richness by limiting individual species' ranges? *Proc. R. Soc. Lond. B* **281,** 1776 (2014).
39. Chao, A. & Jost, L. Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology* **93,** 2533–2547 (2012).
40. Alroy, J. Fair sampling of taxonomic richness and unbiased estimation of origination and extinction rates. *Quantitative Methods in Paleobiology* **16,** 55–80 (2010).
41. Colwell, R. K. & Coddington, J. A. Estimating terrestrial biodiversity through extrapolation. *Phil. Trans. R. Soc. Lond. B* **345,** 101–118 (1994).
42. Oksanen, J. *et al.* The vegan package version 1.15-0. https://cran.r-project.org/web/packages/vegan/index.html (2008).
43. Hsieh, T., Ma, K. & Chao, A. iNEXT online. https://chao.shinyapps.io/iNEXT/ (2016).
44. Kissling, W. D. & Carl, G. Spatial autocorrelation and the selection of simultaneous autoregressive models. *Glob. Ecol. Biogeogr.* **17,** 59–71 (2008).
45. Witman, J. D., Etter, R. J. & Smith, F. The relationship between regional and local species diversity in marine benthic communities: a global perspective. *Proc. Natl Acad. Sci. USA* **101,** 15664–15669 (2004).
46. Bivand, R. *et al.* The spdep package. https://cran.r-project.org/web/packages/spdep/index.html (2016).
47. Dorazio, R. M., Royle, J. A., Söderström, B. & Glimskär, A. Estimating species richness and accumulation by modeling species occurrence and detectability. *Ecology* **87,** 842–854 (2006).
48. Dorazio, R. M. & Royle, J. A. Estimating size and composition of biological communities by modeling the occurrence of species. *J. Am. Stat. Assoc.* **100,** 389–398 (2005).
49. Kéry, M. & Royle, J. Hierarchical Bayes estimation of species richness and occupancy in spatially replicated surveys. *J. Appl. Ecol.* **45,** 589–598 (2008).
50. Gelfand, A. E. *et al.* Modelling species diversity through species level hierarchical modelling. *J. R. Stat. Soc. Ser. C Appl. Stat.* **54,** 1–20 (2005).
51. MacKenzie, D. I. *et al.* Estimating site occupancy rates when detection probabilities are less than one. *Ecology* **83,** 2248–2255 (2002).
52. Plummer, M. in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC)* (2003).
53. Plummer, M. JAGS: Just another Gibbs sampler. http://mcmc-jags.sourceforge.net/ (2004).

© 2016 Macmillan Publishers Limited. All rights reserved

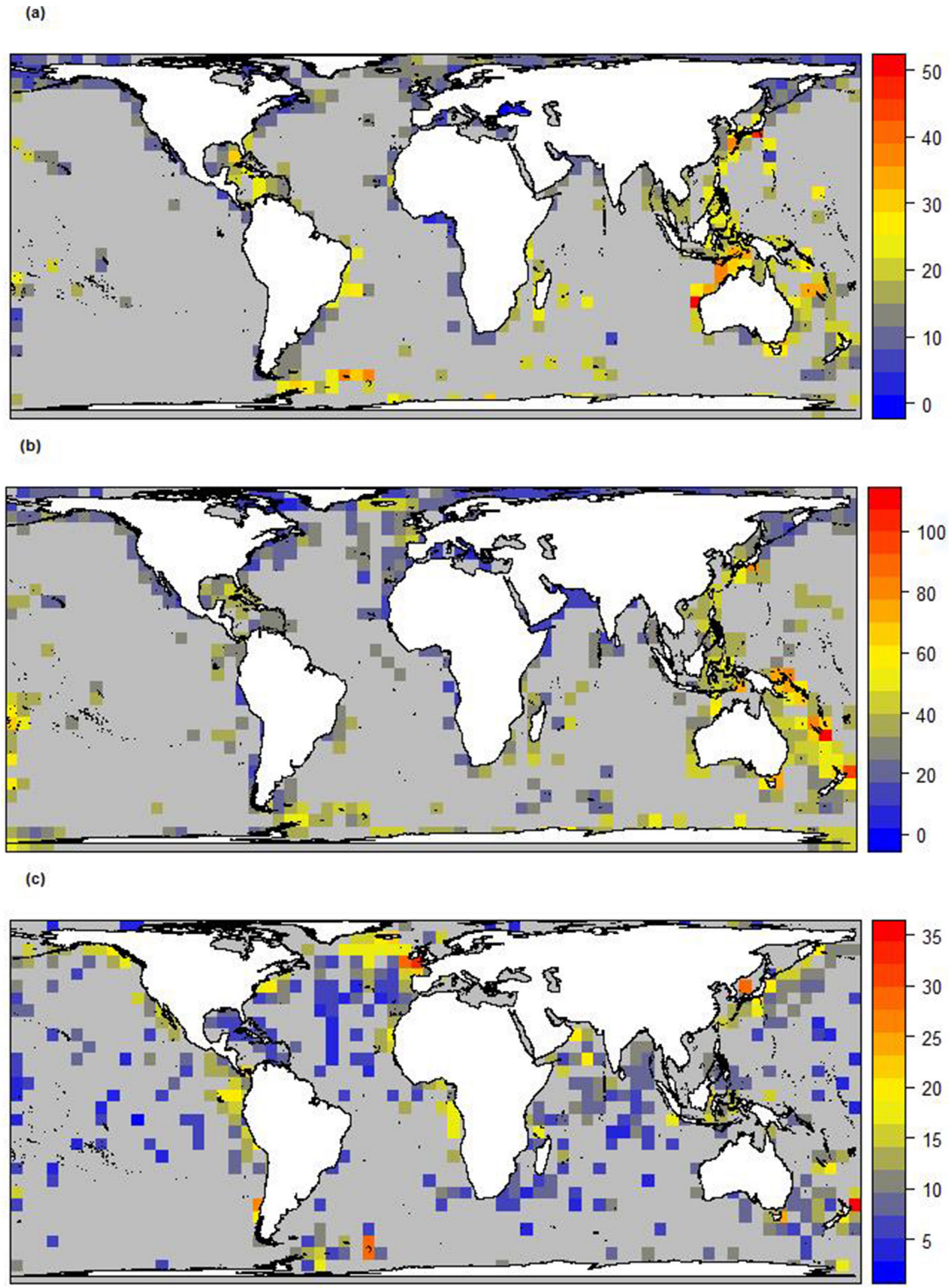
54. Su, Y.-S. & Yajima, M. R2jags: A Package for Running jags from R. http://CRAN.R-project.org/package=R2jags (2012).
55. Gelman, A. & Rubin, D. B. Inference from iterative simulation using multiple sequences. *Stat. Sci.* **7,** 457–472 (1992).
56. Sanderson, C., Curtin, R., Cullinan, I., Bouzas, D. & Funiak, S. Armadillo: C++ linear algebra library. http://arma.sourceforge.net/ (2014).
57. Eddelbuettel, D. *et al.* Rcpp: Seamless R and C++ Integration. http://CRAN.R-Project.org/package=Rcpp (2015).
58. Francois, R., Eddelbuettel, D. & Bates, D. RcppArmadillo: Rcpp integration for Armadillo templated linear algebra library. https://cran.r-project.org/web/packages/RcppArmadillo/ (2012).
59. R Core Team. R: A language and environment for statistical computing. http://www.R-project.org/ (2010).
60. ESRI. 2011. ArcGIS Desktop: Release 10. Redlands, CA: Environmental Systems Research Institute. http://www.esri.com/
61. Keitt, T. H., Bivand, R., Pebesma, E. & Rowlingson, B. rgdal: bindings for the Geospatial Data Abstraction Library. http://CRAN.R-project.org/package=rgdal (2011).
62. Hijmans, R. & van Etten, J. raster: geographic data analysis and modeling. https://cran.r-project.org/web/packages/raster/ (2014).
63. Lewin-Koh, N. J. *et al.* maptools: tools for reading and handling spatial objects. http://CRAN.R-project.org/package=maptools (2011).
64. Hijmans, R. J., Phillips, S., Leathwick, J. & Elith, J. dismo: species distribution modeling. https://cran.r-project.org/web/packages/dismo/ (2012).
65. Vallina, S. M. *et al.* Global relationship between phytoplankton diversity and productivity in the ocean. *Nature Commun.* **5,** 4299 (2014).
66. Gillooly, J. F., Allen, A. P., West, G. B. & Brown, J. H. The rate of DNA evolution: effects of body size and temperature on the molecular clock. *Proc. Natl Acad. Sci. USA* **102,** 140–145 (2005).

© 2016 Macmillan Publishers Limited. All rights reserved



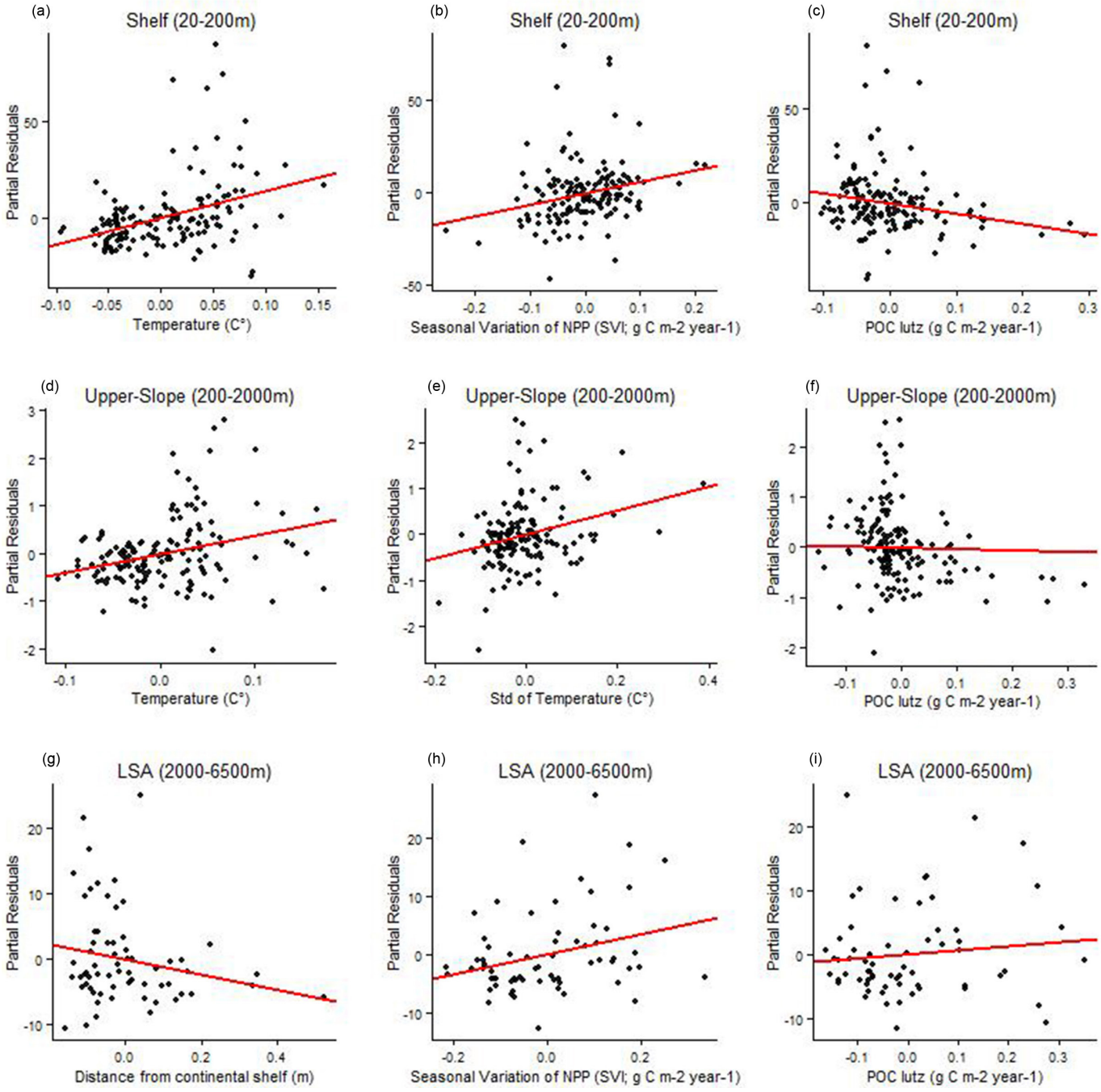
Extended Data Figure 1 | Distribution of global sampling effort across deep-sea bathomes. a–**c**, Spatial plots of sampling effort for ophiuroid occurrence data at the same equal-area grid cells used in MSODM at 500-km equal area grid cells, maximum effort is capped at 100 surveys to help visualize low and high regions of repeated effort. Shelf effort from 20–200 m (**a**), slope effort from 200–2,000 m (**b**) and deep-water collection effort from 2,000–6,500 m (**c**). **d**–**f**, Ophiuroid distribution data are presented for shelf (**d**; red; 20–200 m), slope (**e**; orange; 200–2,000 m) and deep-water (**f**; yellow; 2,000–6,500 m); key represents depth profile. Land maps used in these figures are from the public domain Natural Earth and freely available for personal, educational, and commercial purposes. See http://www.naturalearthdata.com/about/terms-of-use/ for more details.

© 2016 Macmillan Publishers Limited. All rights reserved



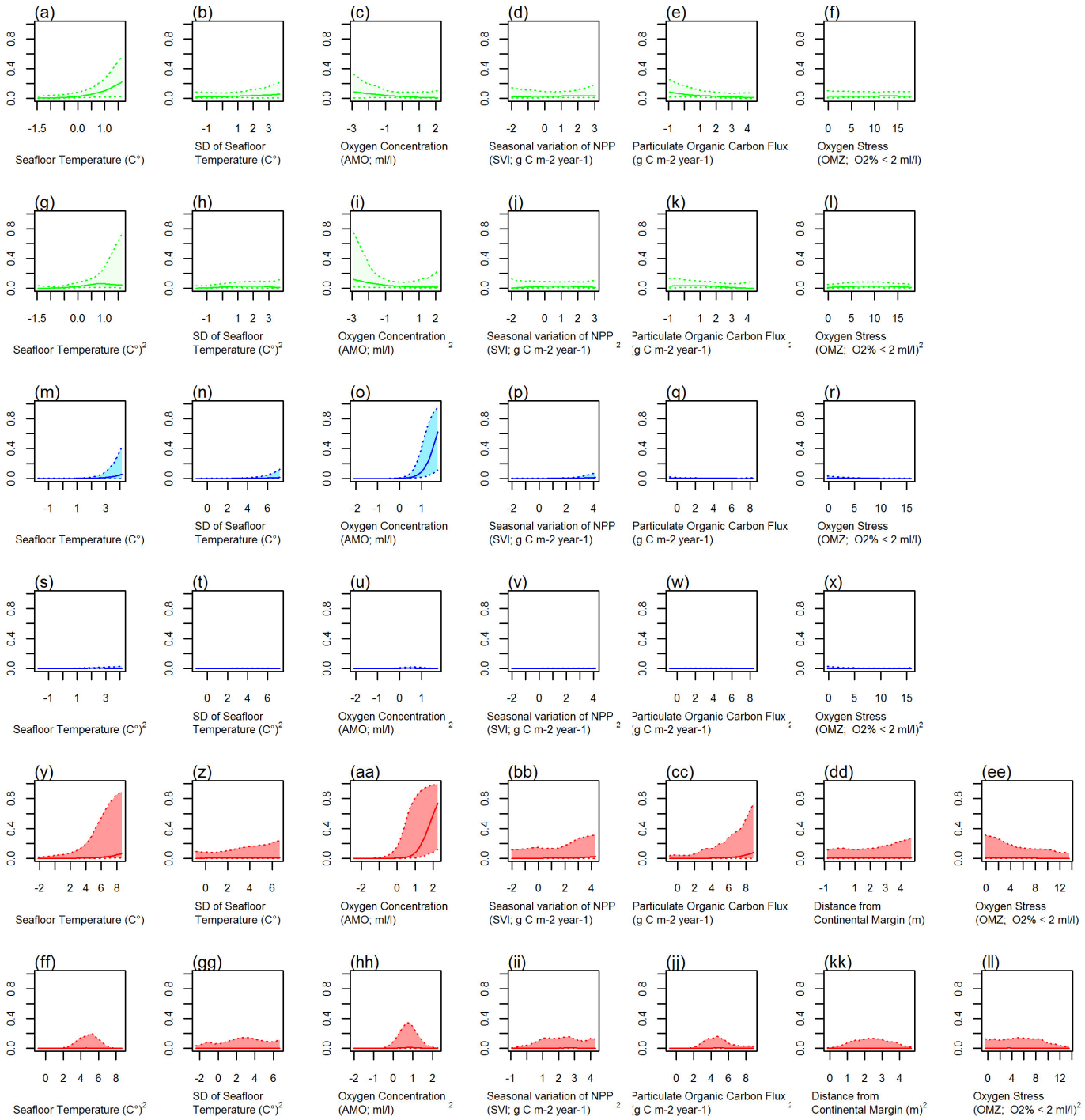
Extended Data Figure 2 | Model estimated global deep-sea species richness across different depth strata. a–c, Maps of species counts ($\hat{N}$) as calculated using MSODM are presented as shelf (**a**), slope (**b**) and deep-water species (**c**). $\hat{N}$ is an estimate of species present per cell based on our occurrence matrix (Z). Z is a latent variable used to calculated presences and absences of species within each cell. The estimated count of species ($\hat{N}$) per 500km equal area grid cell are reported in each figure key. Land maps used in these figures are from the public domain Natural Earth and freely available for personal, educational, and commercial purposes. See http://www.naturalearthdata.com/about/terms-of-use/ for more details.

© 2016 Macmillan Publishers Limited. All rights reserved



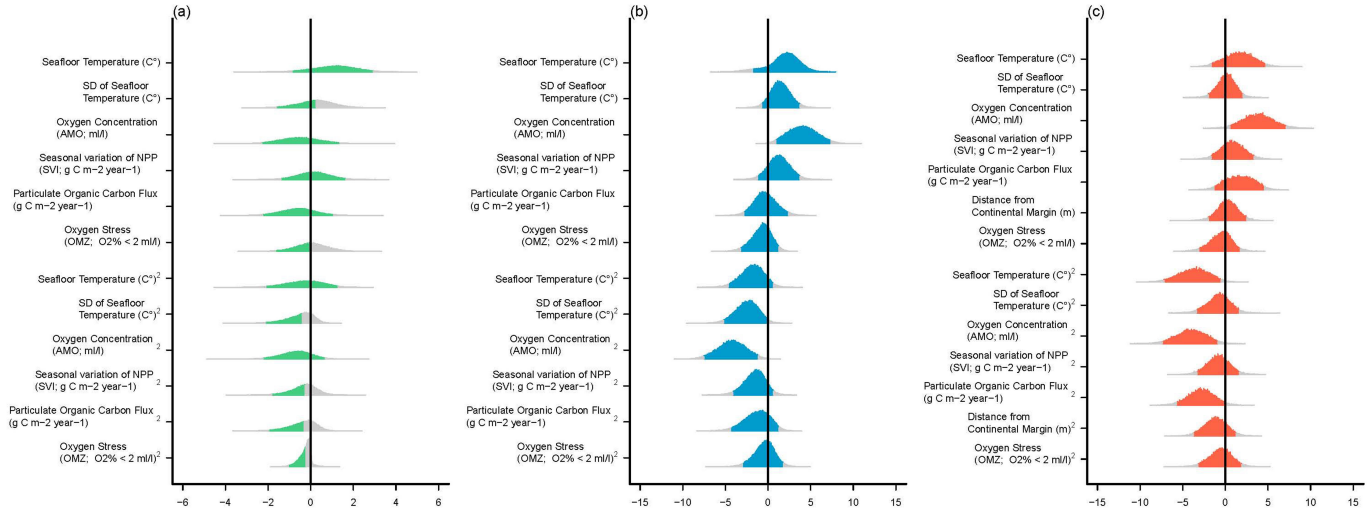
Extended Data Figure 3 | Linear partial residual plots as derived from SLMs. a–i, Partial residual plots for significant variables included in the models for global deep-sea richness at shelf (**a–c**, 20–200 m), upper-slope (**d–f**, 200–2,000 m) and deep-water (LSA; **g–i**, 2,000–6,500 m). Hatched lines are partial fits (red lines). Values on the *x* axis are centred and normalized (mean = 0, variance = 1), as derived from SLMs.

© 2016 Macmillan Publishers Limited. All rights reserved



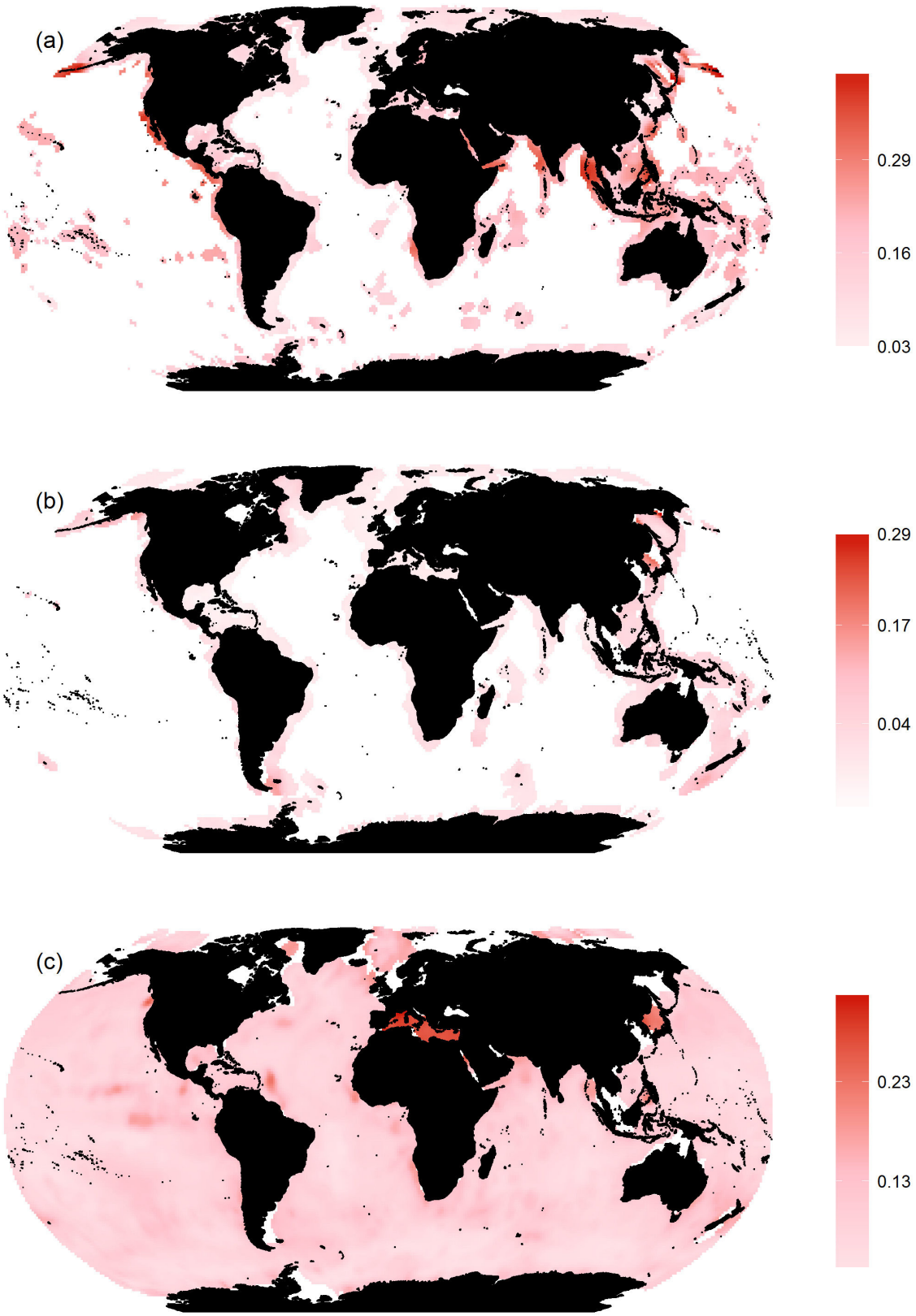
Extended Data Figure 4 | Environmental relationships covariate estimated with the multi-species occupancy–detection model. The shaded areas represent the regions delimited by the tenth–ninetieth percentiles of the estimates obtained from the responses of all species. From top to bottom, rows display the estimates of occupancy (ψ), for shelf (green), slope (blue) and abyss (red) species. All covariates were centred and normalized (mean = 0, variance = 1).

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 5 | Bayesian posterior estimates. a–c, Deep-water MSODM parameter estimates, for shelf (**a**), slope (**b**), and deep-water species (**c**). Posterior distributions of parameter estimates are across all species. All covariates were centred and normalized (mean = 0, variance = 1).

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 6 | Mean variance of MSODM predictions of species occupancy probabilities. Variance for shelf diversity (**a**, 20–200 m), slope diversity (**b**, 200–2,000 m) and deep-water diversity (**c**, 2,000–6,500 m). Land maps used in these figures are from the public domain Natural Earth and freely available for personal, educational, and commercial purposes. See http://www.naturalearthdata.com/about/terms-of-use/ for more details.

© 2016 Macmillan Publishers Limited. All rights reserved

Extended Data Table 1 | Encapsulation of species richness hypotheses by environmental and physical predictors

Hypotheses	Predictors	Predictor Derivation	Hypothesis Description	Citation	Expected pattern anticipated for bathome(s)
Kinetic (thermal) energy Hypothesis	Annual Mean Seafloor Temperature (AMT; C°) Annual Standard Deviation of Seafloor Temperature (ASDT; C°)	CARS 2009 dataset. Interpolated available oceanographic data (from 1950-2009) across the globe for 79 depth layers at a resolution of 0.5° latitude/longitude. averaged	Warmer temperatures lead to higher richness. Possible mechanisms include (but not limited to): (i) increased metabolic rates leading to increased rates of speciation / decreased extinction rates (ii) warmer environments are easier to tolerate physiologically; (iii) warmer environments provide more metabolic niches.	18,12,13	Shelf
Productivity-richness hypothesis	Modelled Particulate Organic Carbon Flux (POC flux; g C m^{-2} $year^{-1}$) Modelled Seasonal variation of net primary productivity (SVI; g C m-2 year-1)	Particulate organic carbon flux to the seafloor (POC flux; g C m-2 year-1) was estimated using NPP and SVI data and productivity export models. SVI is a function of satellite-derived chlorophyll (SeaWiFS). NPP and SVI were calculated across the years 2003 to 2010.	This hypothesis' predicts a positive effect of primary productivity on species richness. Theory and empirical studies also support the idea that seasonal pulses of productivity can sustain diverse populations.	12,19,65	Slope; Deep-Water
The environmental stress hypothesis	Oxygen Minimization Zones (OMZs), Proportion of grid cells with AMO concentrations < 2 ml/l.	CARS 2009 dataset. Interpolated available oceanographic data (from 1950-2009) across the globe for 79 depth layers at a resolution of 0.5° latitude/longitude. averaged	Predicts a negative relationship with environmental stress and richness. Low oxygen concentration is an important environmental stressor in marine systems.	14	Shelf; Slope
Evolutionary rates hypothesis	Annual Mean Oxygen Concentration (AMO; ml/l)	CARS 2009 dataset. Interpolated available oceanographic data (from 1950-2009) across the globe for 79 depth layers at a resolution of 0.5° latitude/longitude. averaged	Oxygen increases mutation rates, but directly contributing to increasing metabolic rates and production of free radicals and oxygen reactive species.	13,66	Shelf, Slope, Deep-Water.
Source-sink Hypothesis	Distance from Coast/Continental Margin	We estimated distance from continental margins using IFREMER Continental margins shape files.	Source-sink hypothesis predicts that species assemblage of abyssal species form a source-sink system in which abyssal populations (sink) are regulated by immigration from bathyal sources.	16,17	Deep-Water

References 12–14, 16–19, 65, 66 are cited in this table.

© 2016 Macmillan Publishers Limited. All rights reserved

Extended Data Table 2 | Correlations between environmental predictors used in GLMs, SLMs and MSODMs by bathome

Bathome Shelf	AMT	ASDT	AMO	NPP	POC SVI	POC	OMZ	DC
AMT	1.000	0.308	-0.780	0.038	-0.228	0.000	0.195	-
ASDT	0.308	1.000	-0.180	0.446	0.171	0.408	0.115	-
AMO	-0.780	-0.180	1.000	-0.237	-0.015	-0.251	-0.492	-
NPP	0.038	0.446	-0.237	1.000	0.357	0.939	0.338	-
POC SVI	-0.228	0.171	-0.015	0.357	1.000	0.591	0.197	-
POC	0.000	0.408	-0.251	0.939	0.591	1.000	0.360	-
OMZ	0.195	0.115	-0.492	0.338	0.197	0.360	1.000	-
DC	-	-	-	-	-	-	-	-
Slope	**AMT**	**ASDT**	**AMO**	**NPP**	**POC SVI**	**POC**	**OMZ**	**DC**
AMT	1.000	0.337	-0.446	0.196	-0.023	0.152	0.214	-
ASDT	0.337	1.000	0.098	0.411	0.347	0.522	0.118	-
AMO	-0.446	0.098	1.000	-0.124	0.093	-0.056	-0.497	-
NPP	0.196	0.411	-0.124	1.000	0.517	0.899	0.220	-
POC SVI	-0.023	0.347	0.093	0.517	1.000	0.599	0.207	-
POC	0.152	0.522	-0.056	0.899	0.599	1.000	0.188	-
OMZ	0.214	0.118	-0.497	0.220	0.207	0.188	1.000	-
DC	-	-	-	-	-	-	-	-
Deep-water (LSA)	**AMT**	**ASDT**	**AMO**	**NPP**	**POC SVI**	**POC**	**OMZ**	**DC**
AMT	1.000	0.158	-0.210	0.139	-0.066	0.150	0.076	-0.112
ASDT	0.158	1.000	-0.210	0.184	-0.036	0.119	0.020	0.078
AMO	-0.210	-0.210	1.000	-0.044	0.182	0.001	-0.341	-0.113
NPP	0.139	0.184	-0.044	1.000	0.521	0.904	0.177	-0.353
POC SVI	-0.066	-0.036	0.182	0.521	1.000	0.520	0.182	-0.393
POC	0.150	0.119	0.001	0.904	0.520	1.000	0.178	-0.375
OMZ	0.076	0.020	-0.341	0.177	0.182	0.178	1.000	-0.149
DC	-0.112	0.078	-0.113	-0.353	-0.393	-0.375	-0.149	1.000

Correlations with an absolute value of greater than 0.7 are highlighted as bold.

© 2016 Macmillan Publishers Limited. All rights reserved

Extended Data Table 3 | Top SLMs based on AIC under all model selection for each bathome (ΔAIC of <2)

Model	SLMs			
Variable Selection	k	AIC	R^2	Moran's P-value
Shelf				
1 + AMT + AMT² + AMO + AMO² + OMZ + OMZ² + SVI + SVI² + POC + POC²	11	2242.145	0.369	0.265
1 + AMT + AMT² + OMZ + OMZ² + SVI + SVI² + POC + POC²	9	2242.351	0.362	0.188
1 + AMT + AMT² + SVI + SVI² + POC + POC²	7	2243.632	0.351	0.204
1 + AMT + AMT² + OMZ + OMZ² + SVI + SVI²	7	2243.855	0.357	0.280
Slope				
1 + AMT + AMT² + POC + POC²	5	2589.723	0.375	0.681
1 + AMT + AMT² + AMO + AMO² + POC + POC²	7	2591.206	0.373	0.658
Deep-Water (LSA)				
1 + SVI + SVI² + POC + POC²	5	806.272	0.205	0.381
1 + AMT + AMT² + POC + POC²	5	806.408	0.164	0.359
1 + POC + POC²	3	806.601	0.135	0.366
1 + SVI + SVI² + POC + POC² + DC + DC²	7	807.072	0.161	0.381
1 + AMT + AMT² + SVI + SVI² + POC + POC²	7	807.554	0.215	0.374
1 + AMT + AMT² + AMO + AMO² + POC + POC²	7	807.633	0.179	0.369

We present model covariates including linear and quadratic terms, the number of parameters (*k*), Akaike information criterion (AIC), R^2 and Moran's *P* value for each model.

© 2016 Macmillan Publishers Limited. All rights reserved

Extended Data Table 4 | Deviance reduction between null multi-species occupancy detection models and fully fitted models

Bathome	Null Deviance	Full Model Deviance	Null pD	Full Model pD	Δ Deviance	Deviance Reduction
Shelf	109583.51 (BCI: 109247.72, 110162.04)	98126.56 (BCI: 97824.14, 98447.66	101050.02	10412.05	11456.93	11.7%
Slope	173525.87 (BCI: 172516.77, 176573.34)	167309.26 (BIC : 167043.31, 167560.76)	847841.31	13329.09	6216.607	3.6%
Deep-water (LSA)	19133.33 (BCI: 18368.61, 20147.84)	13658.34 (BCI: 13480.75, 13842.62)	130153.60	4314.64	5474.98	28.6%

Estimates are presented with Bayesian confidence interval (BCI) for hierarchical multi-species occupancy detection models. *pD* is a Bayesian statistic that measures deviance, it is represented as: $pD = \frac{\text{var(deviance)}}{2}$, which is calculated in JAGS software.

© 2016 Macmillan Publishers Limited. All rights reserved

LETTER

doi:10.1038/nature17995

Local fitness landscape of the green fluorescent protein

Karen S. Sarkisyan[1,2,3,4,5]*, Dmitry A. Bolotin[1,3]*, Margarita V. Meer[4,5], Dinara R. Usmanova[4,5,6], Alexander S. Mishin[1,2], George V. Sharonov[1,7], Dmitry N. Ivankov[4,5,8], Nina G. Bozhanova[1], Mikhail S. Baranov[1,9], Onuralp Soylemez[4,5], Natalya S. Bogatyreva[4,5,8], Peter K. Vlasov[4,5], Evgeny S. Egorov[1], Maria D. Logacheva[9,10,11], Alexey S. Kondrashov[11,12], Dmitry M. Chudakov[1,3], Ekaterina V. Putintseva[1,3], Ilgar Z. Mamedov[1,3], Dan S. Tawfik[13], Konstantin A. Lukyanov[1,2] & Fyodor A. Kondrashov[4,5,14]

Fitness landscapes[1,2] depict how genotypes manifest at the phenotypic level and form the basis of our understanding of many areas of biology[2–7], yet their properties remain elusive. Previous studies have analysed specific genes, often using their function as a proxy for fitness[2,4], experimentally assessing the effect on function of single mutations and their combinations in a specific sequence[2,5,8–15] or in different sequences[2,3,5,16–18]. However, systematic high-throughput studies of the local fitness landscape of an entire protein have not yet been reported. Here we visualize an extensive region of the local fitness landscape of the green fluorescent protein from *Aequorea victoria* (avGFP) by measuring the native function (fluorescence) of tens of thousands of derivative genotypes of avGFP. We show that the fitness landscape of avGFP is narrow, with 3/4 of the derivatives with a single mutation showing reduced fluorescence and half of the derivatives with four mutations being completely non-fluorescent. The narrowness is enhanced by epistasis, which was detected in up to 30% of genotypes with multiple mutations and mostly occurred through the cumulative effect of slightly deleterious mutations causing a threshold-like decrease in protein stability and a concomitant loss of fluorescence. A model of orthologous sequence divergence spanning hundreds of millions of years predicted the extent of epistasis in our data, indicating congruence between the fitness landscape properties at the local and global scales. The characterization of the local fitness landscape of avGFP has important implications for several fields including molecular evolution, population genetics and protein design.

We assayed the local fitness landscape of avGFP by estimating the fluorescence levels of genotypes obtained by random mutagenesis of the avGFP sequence (Fig. 1). We used fluorescence-activated cell sorting (Supplementary Fig. 1) and sequenced the entire GFP coding region to assay the fluorescence of many thousands of genotypes created by random mutagenesis of the wild-type sequence (Supplementary Information 2 and Extended Data Fig. 1). We applied several strategies to minimize the error of our estimate of fluorescence (Supplementary Information 3.4 and 4.4), which was estimated from thousands of independent measurements of the wild-type sequence (false negative error rate = 0.08%) and genotypes incorporating mutations known to eliminate fluorescence (false positive error rate = 0.24%). Our final data set included 56,086 unique nucleotide sequences coding for 51,715 different protein sequences. Our procedure introduced an average of 3.7 mutations per gene sequence, and most assayed genotypes contained several, up to 15, missense mutations. Still, since the total number of possible sequences grows exponentially with the number of mutations, the fraction of sampled sequences was tiny for sequences containing more than two mutations (Extended Data Table 1). We used these data to survey the local fitness landscape of GFP, analysing the effect of single, double and multiple mutations.

The distribution of fitness effects of individual missense mutations was assayed by comparing the distribution of fluorescence of wild-type avGFP amino acid sequences, tagged by different molecular barcodes, and the distribution of fluorescence of sequences carrying a single mutation (Supplementary Information 4.1). We found that at least 75% of mutations had a deleterious effect on fluorescence, including 9.4% of single mutations conferring a more than fivefold decrease in fluorescence, but for many mutations the effect was small (Fig. 2a). Accordingly, genotypes with multiple missense mutations were more likely to have low fluorescence, and most genotypes carrying five or more missense mutations were non-fluorescent (Extended Data Fig. 2). Mutations with a strong effect on fluorescence preferably resided at sites that coded for amino acid residues oriented internally towards the chromophore (Fig. 2b, c), which is consistent with data on other proteins on the preference of deleterious mutations to target buried residues[9,11–13]. The effect of mutations on fluorescence was positively correlated with site conservation (Extended Data Fig. 3a, Spearman's rank correlation coefficient = 0.40, $P = 1.44 \times 10^{-10}$), and mutations with a deleterious impact were less likely to be found in orthologous sequences (Extended Data Fig. 3b) and more likely to be found in sites with a deleterious deletion (Supplementary Fig. 2). Still, ~10% of mutant states conferring a non-fluorescent phenotype were nevertheless fixed in long-term evolution (Extended Data Fig. 3b), and a substantial fraction of genotypes containing only mutations leading to amino acid states from GFP orthologues was non-fluorescent (Supplementary Fig. 3), indicating that epistasis affects the avGFP fitness landscape[16].

Interaction of deleterious mutations can manifest in positive epistasis, when the joint effect of mutations is weaker than their independent contribution, or negative epistasis when the joint effect is stronger (Fig. 3a). Light intensity is perceived in the logarithmic scale by living beings, including jellyfish[19]. Thus, we defined epistasis e as the deviation from additivity of effects of single mutations on the logarithmic

[1]Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Miklukho-Maklaya 16/10, 117997 Moscow, Russia. [2]Nizhny Novgorod State Medical Academy, Minin Sq. 10/1, 603005 Nizhny Novgorod, Russia. [3]Central European Institute of Technology, Masaryk University, Brno 62500, Czech Republic. [4]Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, 88 Dr. Aiguader, 08003 Barcelona, Spain. [5]Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain. [6]Moscow Institute of Physics and Technology, Institutskiy Pereulok 9, g.Dolgoprudny 141700, Russia. [7]Faculty of Medicine, Moscow State University, Lomonosov Avenue 31/5, Moscow 119192, Russia. [8]Laboratory of Protein Physics, Institute of Protein Research of the Russian Academy of Sciences, 4 Institutskaya Str., Pushchino, Moscow Region 142290, Russia. [9]Pirogov Russian National Research Medical University, Ostrovitianov 1, Moscow 117997, Russia. [10]A. A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow 127051, Russia. [11]Department of Bioinformatics and Bioengineering, Moscow State University, Moscow 119234, Russia. [12]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan 48109, USA. [13]Department of Biological Chemistry, Weizmann Institute of Science, Rehovot 76100, Israel. [14]Institució Catalana de Recerca i Estudis Avançats (ICREA), 23 Pg. Lluís Companys, 08010 Barcelona, Spain.
*These authors contributed equally to this work.

© 2016 Macmillan Publishers Limited. All rights reserved

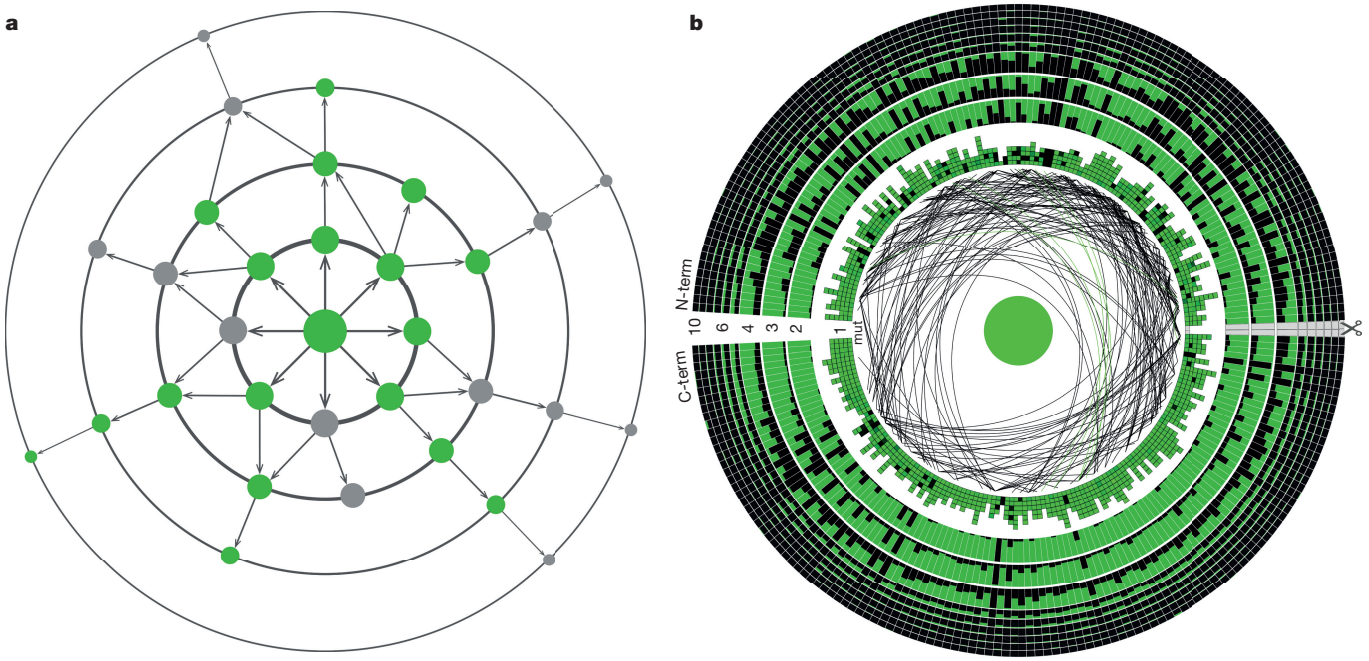


Figure 1 | Exploring the local fitness landscape. a, Wild-type avGFP (centre) and most single mutants (innermost circle) fluoresce green. Genotypes with multiple mutations may exhibit negative epistasis, with combinations of neutral mutations creating non-fluorescent phenotypes (grey), or positive epistasis, in which a mutation in a non-fluorescent genotype restores fluorescence. **b**, The GFP sequence arranged in a circle, each column representing one amino acid site. In the first circle, the colour intensity of the squares indicates the brightness of a single mutation at the corresponding site relative to the wild type, shown in the centre. Sites with positive and negative epistatic interactions between pairs of mutations are connected by green and black lines, respectively. In circles further away from the centre, representing genotypes with multiple mutations, the fraction of the column coloured green (black) represents the fraction of genotypes corresponding to high (low) fluorescence among all assayed genotypes with a mutation at that site. Scissors indicate the restriction site.

scale. We compared the decrement of the log-fluorescence of a multiple mutant F_{mult} to the sum of decrements of individual mutants, such that $e = (F_{\mathrm{mult}} - F_{\mathrm{WT}}) - \sum_i (F_i - F_{\mathrm{WT}})$, where F_{WT} and F_i are the log-fluorescence values conferred by wild-type avGFP and avGFP with the i-th single missense mutation, respectively. We restricted the expected fluorescence of the multiple mutant $\sum_i (F_i - F_{\mathrm{WT}})$ to the observed maximum or minimum levels (Supplementary Information 4.2). This eliminated spurious detection of epistasis, which could otherwise occur, for example, when a non-fluorescent double mutant consists of two mutations, both of which individually confer a non-fluorescent genotype. We defined strong epistasis as $|e| > 0.7$, or as cases in which the observed fluorescence differed from the expected by at least five-fold, with a false discovery rate of $<1\%$ (Supplementary Fig. 4).

Negative epistasis affected up to 30% of all genotypes, depending on the number of mutations (Fig. 3b, c), which resulted in a larger than expected fraction of non-fluorescent genotypes (Fig. 3c). Genotypes

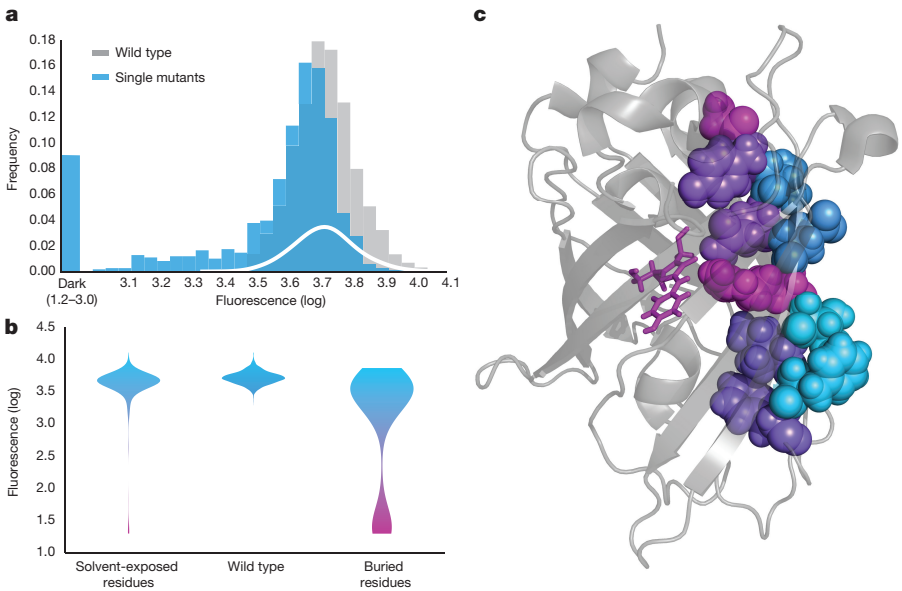


Figure 2 | The effect of single mutations on avGFP. a, The distributions of independently measured fluorescence for 2,442 wild-type sequences (grey), 1,114 single mutants (blue), and the estimated fraction of neutral mutations (white). **b**, **c**, Single missense mutations strongly decreasing fluorescence (violet) tended to occur at sites with internally oriented residues (**b**), shown on a selected β-strand of the GFP structure (**c**).

© 2016 Macmillan Publishers Limited. All rights reserved

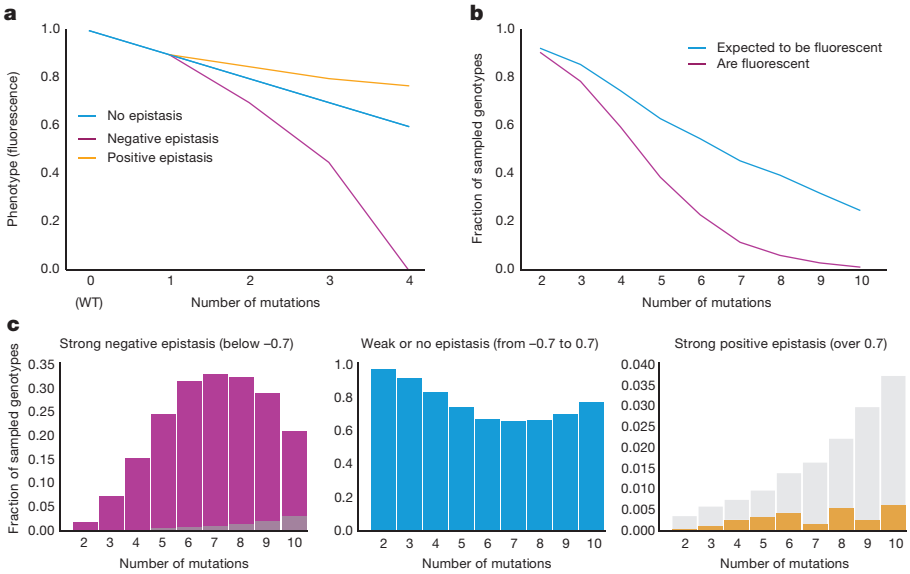


Figure 3 | Prevalence of epistasis in the local fitness landscape of avGFP. **a**, A hypothetical representation of negative and positive epistasis as a function of the number of single mutations from avGFP. WT, wild type. **b**, The fraction of observed non-fluorescent genotypes (red) and the expected fraction of non-fluorescent genotypes calculated as the sum of the log-fluorescence effects of individual mutations (blue). **c**, The distributions of epistasis for negative and positive epistasis of different strength, with the expected false discovery rate shown in grey.

carrying more than seven mutations showed a decrease in the prevalence of negative epistasis because many genotypes carrying multiple mutations were expected to lose fluorescence even without epistasis (Fig. 3b). Positive epistasis was rare in avGFP, on the order of accuracy of our method. We sampled ~2% of all possible pairs of mutations (Extended Data Table 1), assaying 30% of pairs of amino acid sites (16,898 out of 55,696, Extended Data Fig. 4a). Epistatic pairs of sites were located across the avGFP sequence (Extended Data Fig. 4a), mostly beyond the range of direct physical interaction of amino acid residues (Extended Data Fig. 4b) but marginally closer together than random (Extended Data Fig. 4c, $P < 0.0004$, Mann–Whitney U-test). Epistasis was found among 96% of mutations with weak effect (Extended Data Fig. 4d), suggesting that their joint effect brings the protein over some stability margin[8,20]. Finally, epistasis was more common between pairs of sites in which both residues are internally oriented (Extended Data Fig. 4e). Taken together, these data indicate that epistasis was more common at functionally important sites.

In a unidimensional landscape, fitness is a monotonic function of an intermediate variable known as fitness potential[21,22], which is the sum of effects of individual mutations. We used multiple regression considering a non-epistatic fitness function in which log-fluorescence, F, is equal to the linear predictor, the fitness potential, p, such that $F = f(p) = p$. This simplest, non-epistatic model explained only 70% of the initial sample variance ($\sigma^2 = 1.12$ and $\sigma^2 = 0.34$ before and after the application of the model, respectively). Using the variance of the 2,442 wild-type fluorescence measurements, we estimated that ~1% of the initial sample variance can be attributed to noise ($\sigma^2 = 0.0097$), indicating that the remaining 29% of sample variance cannot be explained without epistasis.

The simplest form of an epistatic fitness function is when fitness is a monotonic nonlinear function of p[21,22]. The lack of genotypes with intermediate fluorescence (Extended Data Fig. 5a) suggests that the avGFP fitneses landscape can be described by a truncation-like fitness function[23]. We therefore modelled F as a sigmoid function of p, which explained 85% of the initial sample variance ($\sigma^2 = 0.17$). A more complex sigmoid-shaped fitness function refined with a neural network approach (Supplementary Information 4.6) explained 93.5% of the initial sample variance ($\sigma^2 = 0.065$, Extended Data Fig. 5), confirming that the fitness landscape can mostly be represented by a unidimensional threshold function (Fig. 4), which can arise from the joint contribution of mutations to protein stability[8,13,14,20,24]. The average fluorescence of single mutants of avGFP as a function of the predicted protein destabilization, $\Delta\Delta G$, reveals a threshold around 7–9 kcal mol^{-1} (Fig. 4). Notably, the hidden value found by the artificial neural network for single mutants correlated to the predicted $\Delta\Delta G$ (Fig. 4 and Extended Data Fig. 5f), confirming a probable influence of protein stability on the nature of epistasis in avGFP. The threshold fitness function does a remarkably good job in approximating the entire fitness landscape, explaining ~95% of all variance. However, when taking into account the error rate of our data set, we estimate that at least 0.3% of genotypes cannot be explained by the threshold fitness function (Supplementary Information 4.5 and Extended Data Fig. 5d), representing instances of multidimensional epistasis[2,5,7].

We compared the local avGFP fitness peak to the global GFP fitness landscape using sequences of GFP orthologues. Negative, threshold-like epistasis, leading to truncation selection[23] against slightly

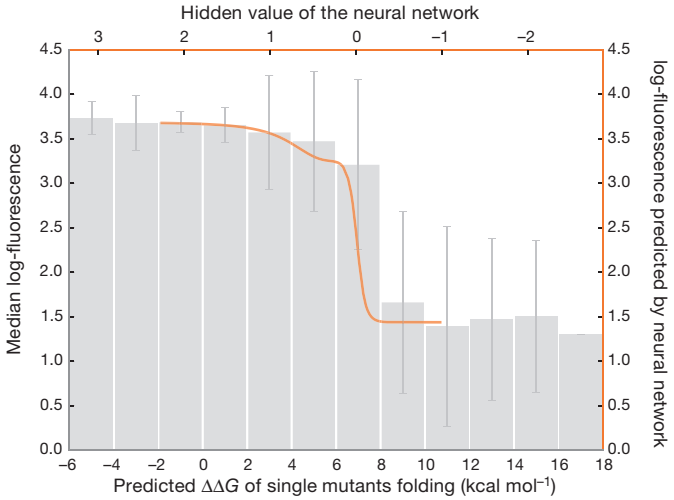


Figure 4 | Modelling genotype to phenotype relationship. Median fluorescence of GFP with single mutations as a function of their effect on predicted folding energy ($\Delta\Delta G$), overlaid with the independently obtained sigmoid-like fitness function predicted by the neural network (orange line). Error bars denote s.d.

© 2016 Macmillan Publishers Limited. All rights reserved

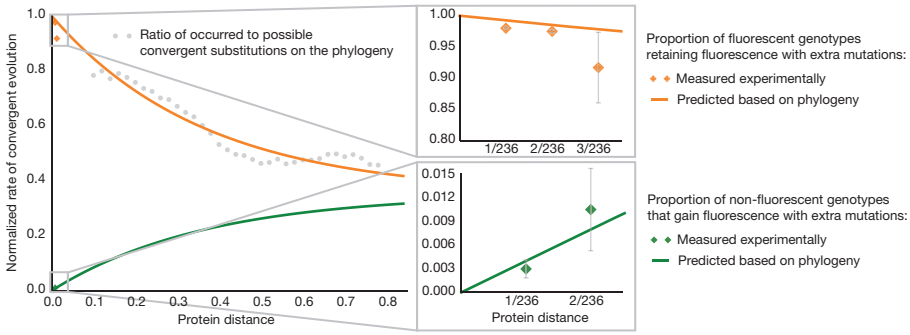


Figure 5 | The fitness matrix model of GFP long-term evolution. The normalized rate of convergent evolution to terminal and reconstructed ancestral amino acid states for each distance bin (grey dots). The expected (orange line) and observed in experimental data (orange dots) probability that a single mutation remains fluorescent as the sequence accumulates other substitutions. The expected (green line) and observed (green dots) probability that a non-fluorescent mutation becomes fluorescent with sequence divergence. Bars represent a binomial proportion confidence interval (confidence level 68%).

deleterious mutations may prevent their accumulation in evolution[25,26]. Thus, we compared the fraction of neutral single mutations to the rate of nonsynonymous and synonymous evolution (d*N*/d*S*), a proxy for the average strength of selection. The average d*N*/d*S* across a broad phylogenetic range was 0.35 ± 0.1 (mean $\pm$ s.d.), and 0.17 when avGFP was compared to the orthologue from the closest fluorescent relative *Aequorea macrodactyla*. These measurements are similar to the estimated proportion of neutral mutations in avGFP (0.23), suggesting that the proportion of neutral mutations is similar across distant fitness peaks. The rate at which the phenotypic effect of amino acid substitutions changes across evolution, which is reflected in the changing rate of convergent evolution across the phylogenetic tree[27,28], can be used to model the prevalence of epistasis with the fitness matrix model[28]. This model approximates the prevalence of epistasis as the proportion of amino acid mutations that markedly change their effect on fitness after the occurrence of substitutions at other sites (Supplementary Information 5). Applying the fitness matrix model to the GFP multiple alignment, we predicted the proportion of mutations that change their effects on fluorescence when found in a different genetic background, revealing prevalence of positive and negative epistasis, which concur with our experimental observations (Fig. 5 and Supplementary Fig. 5). The congruence of the data from an evolutionary trajectory spanning hundreds of millions of years with experimental data are remarkable, suggesting similarity in the local and global structures of the fitness landscape shaped by strong epistatic interactions.

Our study provides complementary results of the analysis of single and double mutations to several previous studies[9,11–14], and a novel depiction of a large segment of the fitness landscape of a single protein. The proportion of neutral single mutations in our data was similar to that observed when fitness was assayed directly or through competition experiments[10,24,29] but substantially lower than that observed in functional studies[4,10,11,13,17]. Furthermore, the propensity of multiple mutations to have a stronger negative effect on fitness than the sum of individual mutation effects has been observed[9,10,12,14]. However, because our analysis considered genotypes carrying multiple mutations, we infer a wider picture of the local fitness landscape (Supplementary Video 1). The avGFP fitness peak is narrow and defined by negative epistasis, best described by truncation selection in which fluorescence is eliminated if the joint effect of mutations exceeds a threshold of an intermediate property, possibly protein stability[8,20]. Such a landscape increases the efficacy of selection against slightly deleterious mutations[23], preventing their accumulation in evolution[25,26]. Simultaneously, the fitness landscape is approximately non-epistatic near the fitness peak. If other proteins have a similar fitness landscape it would support the nearly neutral theory of evolution[30], explaining the selective forces and evolutionary dynamics of mutations with negligible individual effects on fitness.

The broad congruence of our data with the prevalence of epistasis from long-term evolution suggests that the shape of the local fitness landscape can be extrapolated to a larger scale. Yet, epistasis between sites coding for residues with a direct interaction in protein structure was rare, contrasting with observation of such instances in long-term evolution[16] and a mutation assay of the RNA recognition motif (RRM) domain[12]. Thus, the local fitness landscape spanning a few mutations from a single fitness peak may be approximated by a unidimensional threshold fitness potential function; however, this simple fitness function may not be appropriate to describe fitness landscapes that incorporate fitness ridges connecting sequences of more divergent orthologues[27]. The nature of global fitness landscapes, especially the interaction between local and global scales, remains to be explored.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 28 April 2015; accepted 7 April 2016.
Published online 11 May 2016.

1. Wright, S. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proc. Sixth Int. Congr. Genet.* **1,** 356–366 (1932).
2. de Visser, J. A. G. M. & Krug, J. Empirical fitness landscapes and the predictability of evolution. *Nature Rev. Genet.* **15,** 480–490 (2014).
3. Dean, A. M. & Thornton, J. W. Mechanistic approaches to the study of evolution: the functional synthesis. *Nature Rev. Genet.* **8,** 675–688 (2007).
4. Soskine, M. & Tawfik, D. S. Mutational effects and the evolution of new protein functions. *Nature Rev. Genet.* **11,** 572–582 (2010).
5. Weinreich, D. M., Lan, Y., Wylie, C. S. & Heckendorn, R. B. Should evolutionary geneticists worry about higher-order epistasis? *Curr. Opin. Genet. Dev.* **23,** 700–707 (2013).
6. Mackay, T. F. C. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nature Rev. Genet.* **15,** 22–33 (2014).
7. Taylor, M. B. & Ehrenreich, I. M. Higher-order genetic interactions and their contribution to complex traits. *Trends Genet.* **31,** 34–40 (2015).
8. Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N. & Tawfik, D. S. Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* **444,** 929–932 (2006).
9. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nature Methods* **7,** 741–746 (2010).
10. Roscoe, B. P., Thayer, K. M., Zeldovich, K. B., Fushman, D. & Bolon, D. N. Analyses of the effects of all ubiquitin point mutants on yeast growth rate. *J. Mol. Biol.* **425,** 1363–1377 (2013).
11. Jacquier, H. *et al.* Capturing the mutational landscape of the beta-lactamase TEM-1. *Proc. Natl Acad. Sci. USA* **110,** 13067–13072 (2013).
12. Melamed, D., Young, D. L., Gamble, C. E., Miller, C. R. & Fields, S. Deep mutational scanning of an RRM domain of the *Saccharomyces cerevisiae* poly(A)-binding protein. *RNA* **19,** 1537–1551 (2013).
13. Olson, C. A., Wu, N. C. & Sun, R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr. Biol.* **24,** 2643–2651 (2014).
14. Bank, C., Hietpas, R. T., Jensen, J. D. & Bolon, D. N. A systematic survey of an intragenic epistatic landscape. *Mol. Biol. Evol.* **32,** 229–238 (2015).
15. Meini, M. R., Tomatis, P. E., Weinreich, D. M. & Vila, A. J. Quantitative description of a protein fitness landscape based on molecular features. *Mol. Biol. Evol.* **32,** 1774–1787 (2015).

© 2016 Macmillan Publishers Limited. All rights reserved

16. Kondrashov, A. S., Sunyaev, S. & Kondrashov, F. A. Dobzhansky–Muller incompatibilities in protein evolution. *Proc. Natl Acad. Sci. USA* **99,** 14878–14883 (2002).
17. Firnberg, E., Labonte, J. W., Gray, J. J. & Ostermeier, M. A comprehensive, high-resolution map of a gene's fitness landscape. *Mol. Biol. Evol.* **31,** 1581–1592 (2014).
18. Parera, M. & Martinez, M. A. Strong epistatic interactions within a single protein. *Mol. Biol. Evol.* **31,** 1546–1553 (2014).
19. Coates, M. M., Garm, A., Theobald, J. C., Thompson, S. H. & Nilsson, D. E. The spectral sensitivity of the lens eyes of a box jellyfish, *Tripedalia cystophora* (Conant). *J. Exp. Biol.* **209,** 3758–3765 (2006).
20. DePristo, M. A., Weinreich, D. M. & Hartl, D. L. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nature Rev. Genet.* **6,** 678–687 (2005).
21. Milkman, R. Selection differentials and selection coefficients. *Genetics* **88,** 391–403 (1978).
22. Kimura, M. & Crow, J. F. Effect of overall phenotypic selection on genetic change at individual loci. *Proc. Natl Acad. Sci. USA* **75,** 6168–6171 (1978).
23. Crow, J. F. & Kimura, M. Efficiency of truncation selection. *Proc. Natl Acad. Sci. USA* **76,** 396–399 (1979).
24. Rockah-Shmuel, L., Tóth-Petróczy, Á. & Tawfik, D. S. Systematic mapping of protein mutational space by prolonged drift reveals the deleterious effects of seemingly neutral mutations. *PLOS Comput. Biol.* **11,** e1004421 (2015).
25. Li, W. H. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *J. Mol. Evol.* **24,** 337–345 (1987).
26. Akashi, H. Inferring weak selection from patterns of polymorphism and divergence at 'silent' sites in *Drosophila* DNA. *Genetics* **139,** 1067–1076 (1995).
27. Povolotskaya, I. S. & Kondrashov, F. A. Sequence space and the ongoing expansion of the protein universe. *Nature* **465,** 922–926 (2010).
28. Usmanova, D. R., Ferretti, L., Povolotskaya, I. S., Vlasov, P. K. & Kondrashov, F. A. A model of substitution trajectories in sequence space and long-term protein evolution. *Mol. Biol. Evol.* **32,** 542–554 (2015).
29. Eyre-Walker, A. & Keightley, P. D. The distribution of fitness effects of new mutations. *Nature Rev. Genet.* **8,** 610–618 (2007).
30. Ohta, T. Slightly deleterious mutant substitutions in evolution. *Nature* **246,** 96–98 (1973).

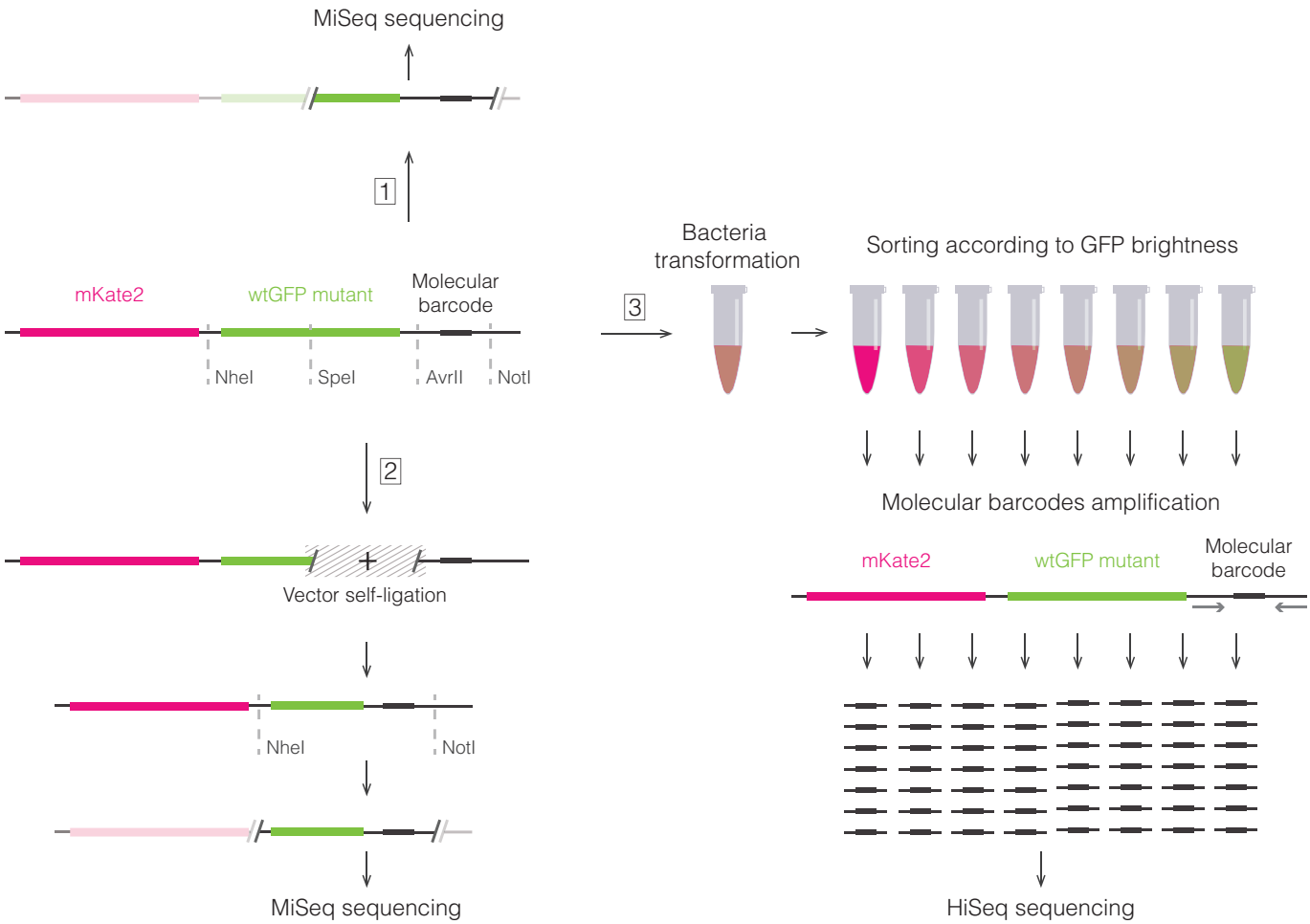
Supplementary Information is available in the online version of the paper.

Acknowledgements We thank Y. Kulikova and G. Filion for discussion on statistical analysis and I. Osterman, R. Moretti and J. Meiler for technical assistance and M. Friesen for a critical reading of the manuscript. We thank H. Himmelbauer, CRG Genomic Unit and the Russian Science Foundation project (14-50-00150) for sequencing. Experiments were partially carried out using the equipment provided by the IBCH core facility (CKP IBCH). The work was supported by HHMI International Early Career Scientist Program (55007424), the EMBO Young Investigator Programme, MINECO (BFU2012-31329), Spanish Ministry of Economy and Competitiveness Centro de Excelencia Severo Ochoa 2013-2017 grant (SEV-2012-0208), Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat's AGAUR program (2014 SGR 0974), Russian Science Foundation (14-25-00129) and the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013, ERC grant agreement, 335980_EinME).

Author Contributions K.S.S. and M.V.M. conceived the idea for the experiment; K.S.S., D.A.B., M.V.M., A.S.M., G.V.S., M.D.L., D.M.C., E.V.P., I.Z.M., D.S.T., K.A.L. and F.A.K. participated in experimental design; K.S.S., D.A.B., M.V.M., G.V.S., E.V.P., E.S.E. and M.D.L. performed the experiments; K.S.S., D.A.B., M.V.M., D.R.U., A.S.M., D.N.I., N.G.B., M.S.B., O.S., N.S.B., P.K.V., A.S.K. and F.A.K. performed data analysis; K.S.S., D.A.B., M.V.M., D.R.U., D.N.I. and F.A.K. wrote the paper.

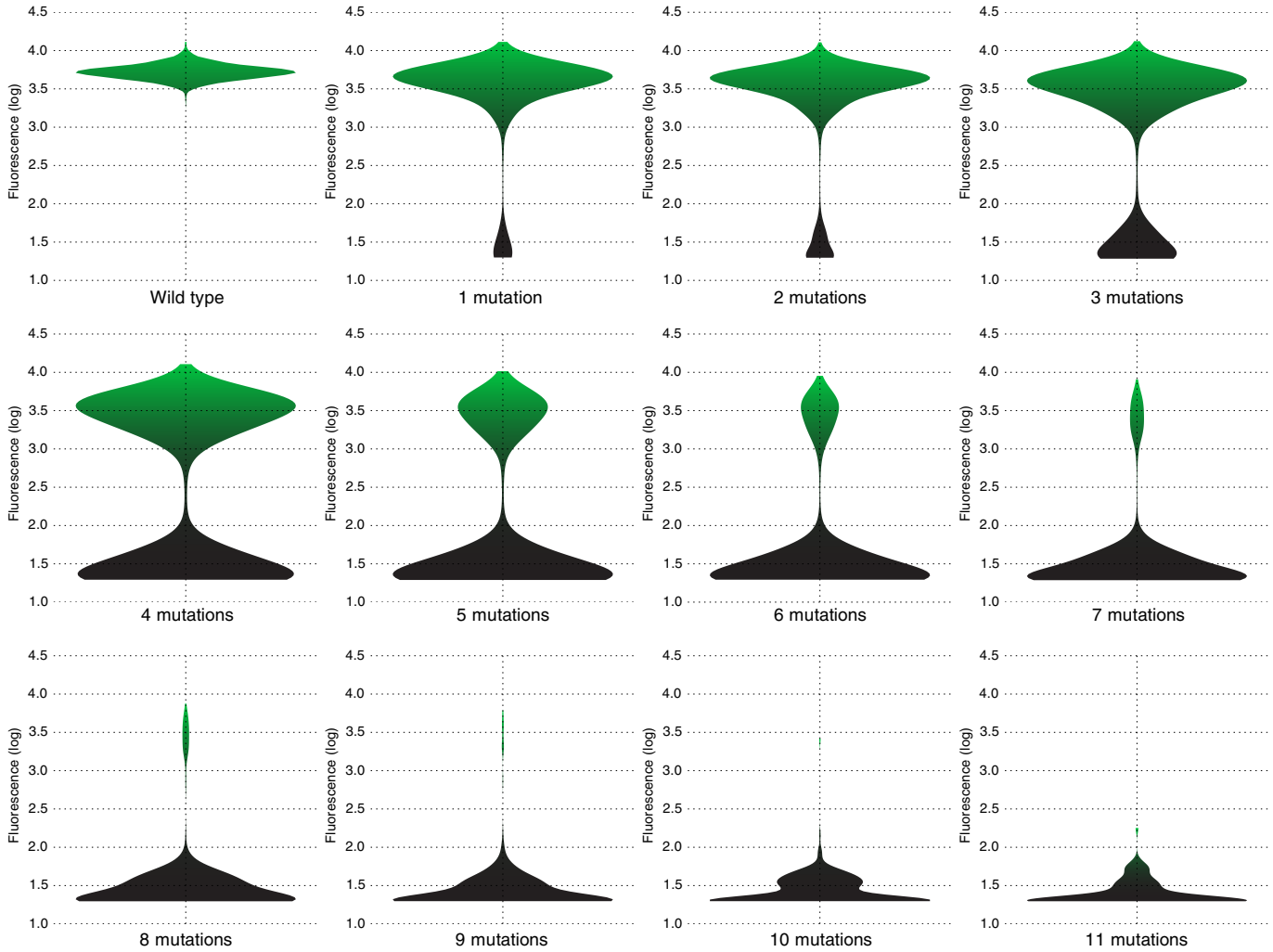
Author Information Raw sequencing data were deposited in the Sequence Read Archive (SRA) under BioProject number PRJNA282342. Processed data sets are available at Figshare http://dx.doi.org/10.6084/m9.figshare.3102154. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to F.A.K. (fyodor.kondrashov@crg.es).

© 2016 Macmillan Publishers Limited. All rights reserved



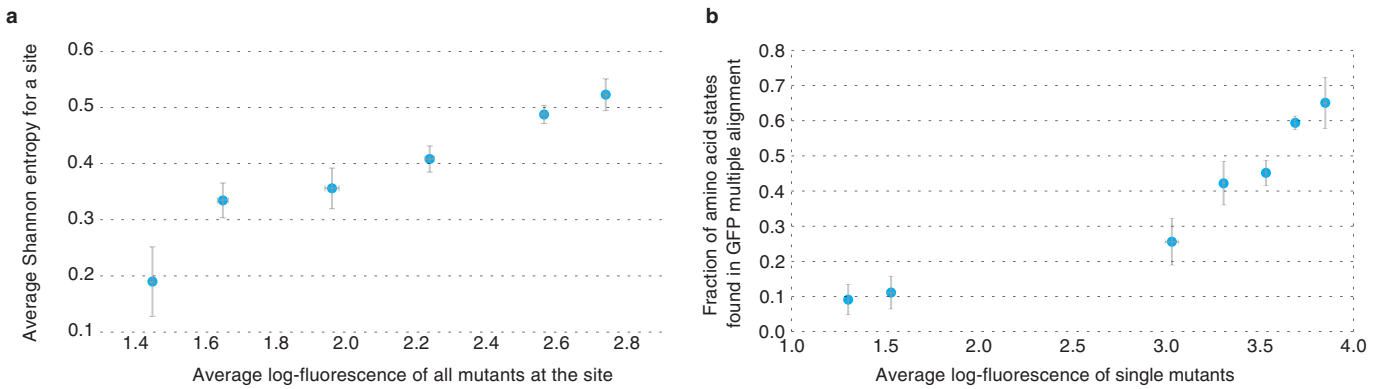
Extended Data Figure 1 | Scheme of the experimental approach. The depiction of the construct design, expression and cell sorting.

© 2016 Macmillan Publishers Limited. All rights reserved



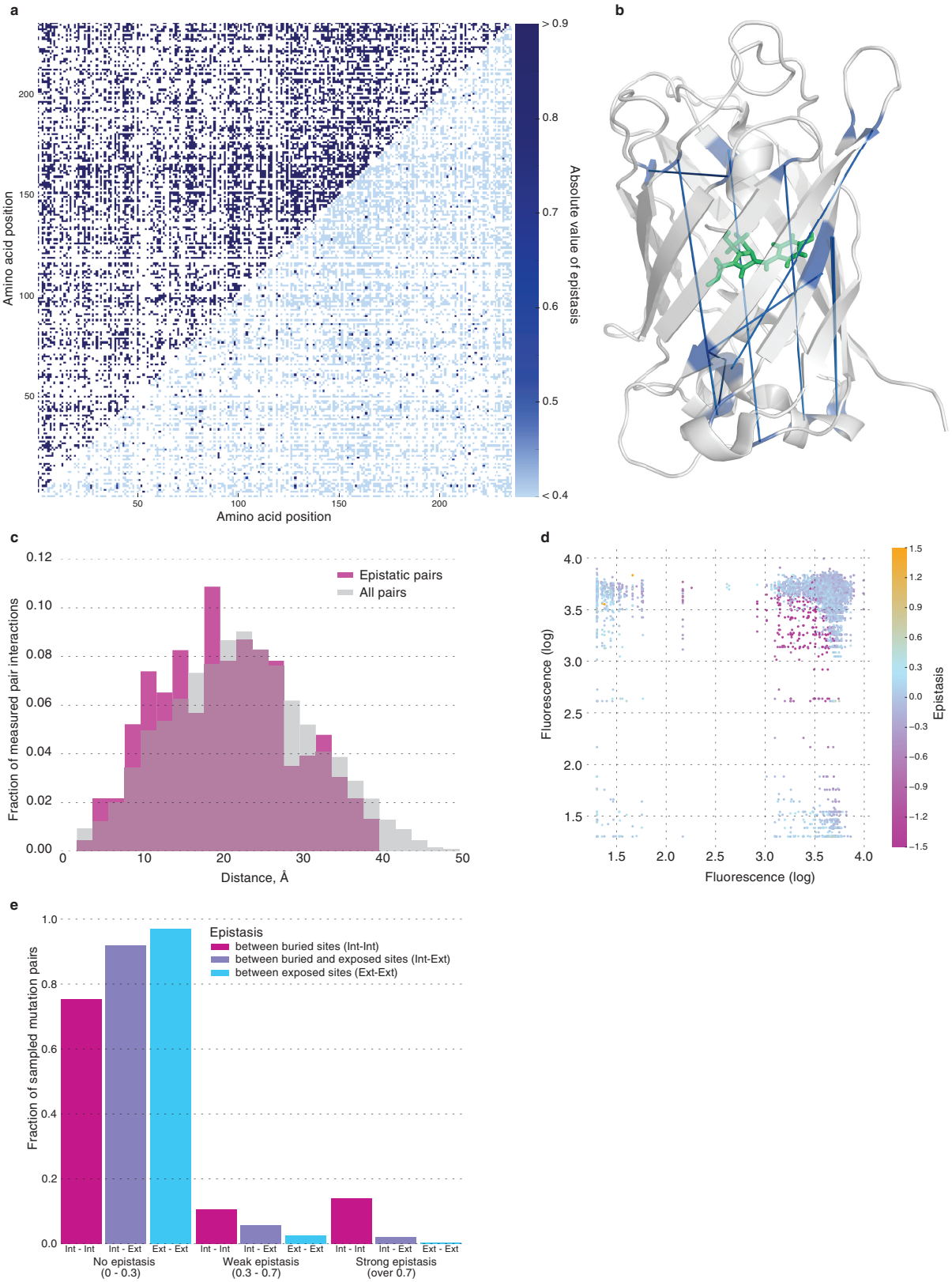
Extended Data Figure 2 | Fluorescence and impact of mutations. A violin plot of the measured levels of fluorescence for genotypes carrying different numbers of missense mutations.

© 2016 Macmillan Publishers Limited. All rights reserved



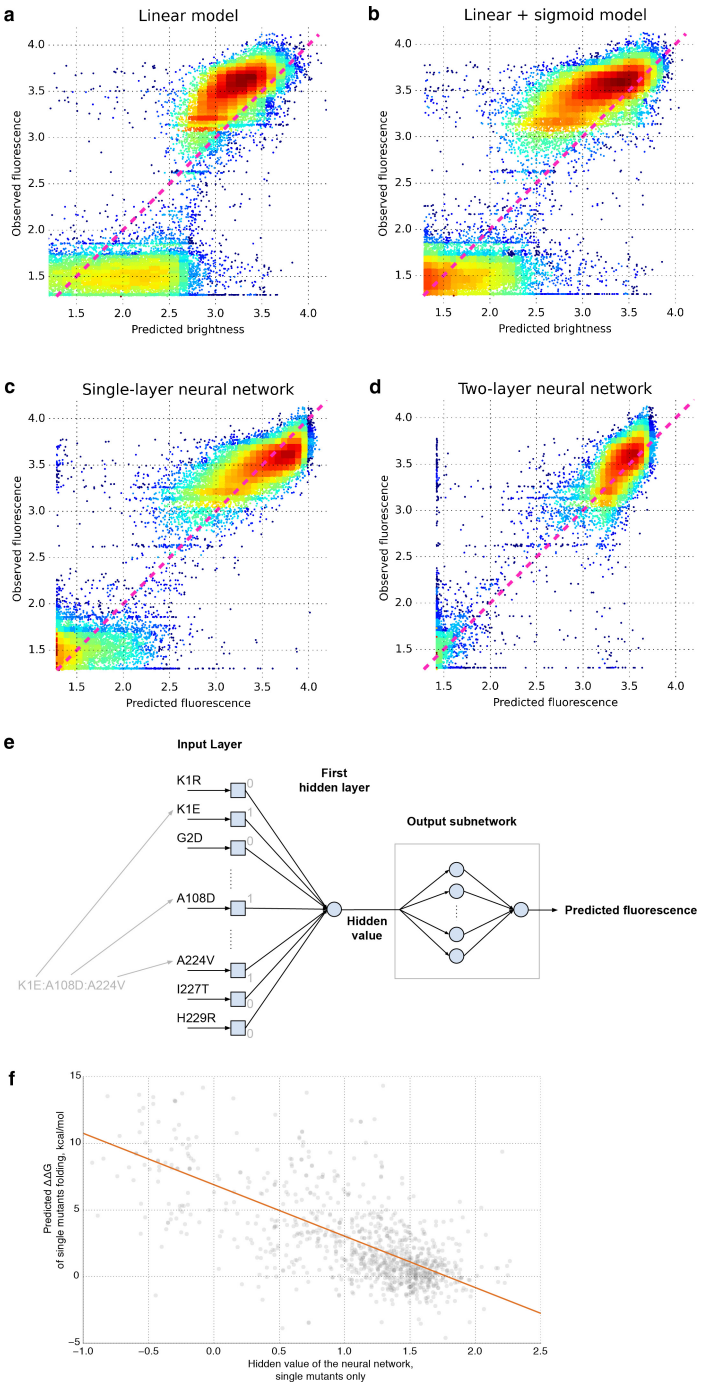
Extended Data Figure 3 | Mutant genotypes and evolution. a, **b**, The log-fluorescence and evolutionary conservation expressed as Shannon entropy (**a**), and fraction of mutant amino acid states found in avGFP orthologues (**b**). The *y*-axis error bars in **b** show the binomial proportion confidence interval level (68%), and other error bars denote s.e.m.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 4 | Epistatically interacting pairs of sites in the GFP structure. a, Pairs of amino acid sites for which we assayed at least one combination of mutations (in blue, top left). The distribution of the maximum level of epistasis observed between sites (blue scale, bottom right) and unknown values (white). **b**, Pairs of sites under exceptionally strong epistatic interaction ($e < -2$) connected by a blue line on the GFP structure. **c**, The distribution of distances in the GFP structure between sites with at least one pair of epistatically interacting mutations (red) and all pairs of sites in the structure (grey). **d**, Epistasis between pairs of mutations as a function of their individual fluorescence. **e**, The contribution of internally and externally oriented amino acid residues in the avGFP structure relative to pairs of missense mutations showing no epistasis ($|e| < 0.3$), weak ($0.3 < |e| < 0.7$) and strong ($|e| > 0.7$) epistasis.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 5 | Modelling effect of mutations on fluorescence. **a**, A multiple linear regression in which fluorescence is linear combination of effects of individual single mutations. **b**, A multiple regression in which mutations contribute linearly to a fitness potential and fluorescence is a sigmoidal function of p where $F \approx e^{-p}/(1+e^{-p})$. **c**, **d**, The predicted fluorescence by a neural network approach. Predicted fitness function by a neural network with one hidden neuron and two neurons in the outer layer. **e**, The scheme of our neural network approach. The genotype data was passed to the input layer of the neural network as an array of 0s or 1s corresponding to the absence or presence of amino acid mutations in the genotype, respectively. The first hidden layer consisted of a single neuron that calculated the weighted sum of inputs using weights obtained during training. The output of the first hidden layer was passed through an output subnetwork that transformed this value with a nonlinear function to make the final prediction of fluorescence. The output subnetwork consisted of several neurons with a sigmoidal transfer function, allowing the subnetwork to approximate a broad range of nonlinear functions. The final mapping of the hidden value to fluorescence was determined by the weights of connections between neurons inside the output subnetwork. During training all weights were optimized to find the best prediction of fluorescence from the hidden value. The resulting function that was defined during training is shown in Fig. 4. **f**, Correlation between the hidden value of the neural network and Rosetta-predicted $\Delta\Delta G$ for single mutants.

© 2016 Macmillan Publishers Limited. All rights reserved

Extended Data Table 1 | Genotypes with measured fluorescence in our data set

Number of mutations from the wild type	Number of amino acid sequences assayed	Number of possible amino acid sequences*	Fraction of amino acid sequences sampled	Number of nucleotide sequences assayed	Number of possible nucleotide sequences	Fraction of nucleotide sequences sampled
1	1,114	1,233	0.90	1,336	2,133	0.626
2	13,010	759,528	0.02	11,555	2,271,645	0.005
3	12,683	311,659,656	4.1×10^{-5}	12,654	1.61×10^{9}	7.86×10^{-6}
4	9,759	9.6×10^{10}	1.0×10^{-7}	10,633	8.55×10^{11}	1.24×10^{-8}
5	7,215	2.4×10^{13}	3.1×10^{-10}	8,164	3.63×10^{14}	2.25×10^{-11}
6	4,643	4.8×10^{15}	9.6×10^{-13}	5,869	1.28×10^{17}	4.58×10^{-14}
7	2,783	8.5×10^{17}	3.3×10^{-15}	3,664	3.87×10^{19}	9.47×10^{-17}
8	1,526	1.3×10^{20}	1.2×10^{-17}	2,212	1.02×10^{22}	2.17×10^{-19}
9	714	1.8×10^{22}	4.1×10^{-20}	1,229	2.39×10^{24}	5.13×10^{-22}
10	352	2.2×10^{24}	1.6×10^{-22}	624	5.04×10^{26}	1.24×10^{-24}

*We considered possible amino acid sequences as sequences in which every amino acid sequence is the result of single nucleotide substitutions from the wild-type avGFP sequence.

© 2016 Macmillan Publishers Limited. All rights reserved

LETTER

doi:10.1038/nature17639

Opponent and bidirectional control of movement velocity in the basal ganglia

Eric A. Yttri[1] & Joshua T. Dudman[1]

For goal-directed behaviour it is critical that we can both select the appropriate action and learn to modify the underlying movements (for example, the pitch of a note or velocity of a reach) to improve outcomes. The basal ganglia are a critical nexus where circuits necessary for the production of behaviour, such as the neocortex and thalamus, are integrated with reward signalling[1] to reinforce successful, purposive actions[2]. The dorsal striatum, a major input structure of basal ganglia, is composed of two opponent pathways, direct and indirect, thought to select actions that elicit positive outcomes and suppress actions that do not, respectively[3,4]. Activity-dependent plasticity modulated by reward is thought to be sufficient for selecting actions in the striatum[5,6]. Although perturbations of basal ganglia function produce profound changes in movement[7], it remains unknown whether activity-dependent plasticity is sufficient to produce learned changes in movement kinematics, such as velocity. Here we use cell-type-specific stimulation in mice delivered in closed loop during movement to demonstrate that activity in either the direct or indirect pathway is sufficient to produce specific and sustained increases or decreases in velocity, without affecting action selection or motivation. These behavioural changes were a form of learning that accumulated over trials, persisted after the cessation of stimulation, and were abolished in the presence of dopamine antagonists. Our results reveal that the direct and indirect pathways can each bidirectionally control movement velocity, demonstrating unprecedented specificity and flexibility in the control of volition by the basal ganglia.

Purposive action requires selection of a goal (for example, go left) and execution parameters (for example, how fast to go). For example, in bird song selection of both discrete, sequential actions (syllables) as well as the pitch can be controlled by reinforcement in cortico-basal ganglia pathways[8,9]. The striatum is a major input nucleus in basal ganglia[1] and the direct and indirect pathway are primarily composed of two molecularly distinct[10] populations of medium spiny projection neurons (MSNs): direct striatonigral (dMSN) and indirect striatopallidal (iMSN) neurons. Sustained activation of dMSNs increases movement, whereas sustained activation of iMSNs reduces movement[11]. As a result, the balance of activity-dependent plasticity at cortical synapses onto dMSNs and iMSNs is thought to underlie the selection of successful goal-directed actions[3,5,12]. While it is known that stimulation of direct pathway neurons can support self-stimulation[13] and bias concomitant choice behaviour[14], there is little direct evidence that MSN activity is sufficient to produce persistent, specific changes in subsequent actions.

We trained mice expressing channelrhodopsin-2 (ChR2) in either dMSNs or iMSNs to perform self-paced, bimanual forelimb movements while head-fixed to obtain a water reward (Fig. 1a; Supplementary Videos 1–4). These single, discrete movements provided a reliable, repeatable behaviour from which we could extract movement parameters (Fig. 1b–d). To determine whether activity in MSNs during a voluntary action is sufficient to control movement parameters, we administered closed-loop photostimulation to the dorsomedial striatum during the fastest third of movements. Stimulation intensity was adjusted to be subthreshold for direct effects on movement, but sufficient to modulate activity to a similar magnitude as endogenous modulation of striatal activity during limb movements (Fig. 1e, f and Extended Data Fig. 1). Stimulation onset occurred within 15 ms of the beginning of a movement and persisted for 450 ms (comparable to movement duration; 505 ms; Fig. 1c, d). To maintain motivation to perform the task independent of stimulation, all movements that crossed the criterion amplitude threshold elicited a delayed liquid reward.

We first asked whether photostimulation of dMSNs during the fastest third of movements could alter the velocity of subsequent movements. Indeed, brief dMSN stimulation was sufficient to produce a significant increase in the peak velocity ($1.4\,\mathrm{cm\,s^{-1}}$ increase from $29.7\,\mathrm{cm\,s^{-1}}$; $P < 7 \times 10^{-5}$; Fig. 2 and Extended Data Fig. 2) of all limb movements. Other movement parameters that were not targeted for closed-loop stimulation such as the amplitude, duration, and tortuosity remained unaltered ($P > 0.7$). This is despite the fact that mice were capable of rapidly adjusting movement parameters to changing reward contingencies (Extended Data Fig. 3). By contrast, iMSNs stimulation during the fastest third of limb movements produced a significant reduction in peak velocity ($-1.1\,\mathrm{cm\,s^{-1}}$; $P < 7 \times 10^{-4}$). The effect of iMSN stimulation had its maximal effect on velocity; movement duration and tortuosity were not significantly altered ($P > 0.3$). Prolonged tonic activation of dMSNs tends to be pro-kinetic in that it evokes generalized increases in voluntary movement ('response vigor'[15]), whereas tonic activation of iMSNs tends to decrease voluntary movement[11]. However, we found that neither brief dMSN nor iMSN stimulation during the fastest movements produced a change in the rate of trial initiation or the rate of licking during reward anticipation and consumption (Fig. 2b and Extended Data Table 1). These results thus demonstrate that closed-loop activation of MSNs is sufficient to produce sustained changes in movement parameters without generalized changes in movement or motivation.

We next examined the effect of successive stimulation on limb movement velocity. If stimulation merely altered the velocity of the current movement, then repeated stimulation should produce an immediate, but constant effect. However, stimulation drove a steady change in velocity that accumulated over the course of several trials (Fig. 2d), apparent in individual sessions (Fig. 2a and Extended Data Fig. 2). We also found that unstimulated movements (trials with subthreshold velocity) were changed to a similar extent. dMSN stimulation produced a $0.9\,\mathrm{cm\,s^{-1}}$ increase ($P = 0.014$) in velocity on unstimulated movements whereas iMSN stimulation produced a $-1.0\,\mathrm{cm\,s^{-1}}$ decrease ($P = 0.001$) in the velocity of unstimulated movements. Moreover, there was no change in variance of the distribution of velocities throughout the session (F-test, $P > 0.5$ for both groups, Extended Data Fig. 4). Together these observations argue that selective stimulation produced a gradual, accumulating shift in the entire distribution of velocities, rather than a change restricted to the stimulated subset (for example, making only fast, stimulated limb movements even faster). These cumulative changes in behaviour may be contrasted with previous reports of optogenetic stimulation that have observed transient effects

[1]Janelia Research Campus, Howard Hughes Medical Institute, 19700 Helix Drive, Ashburn, Virginia 20147, USA.

© 2016 Macmillan Publishers Limited. All rights reserved

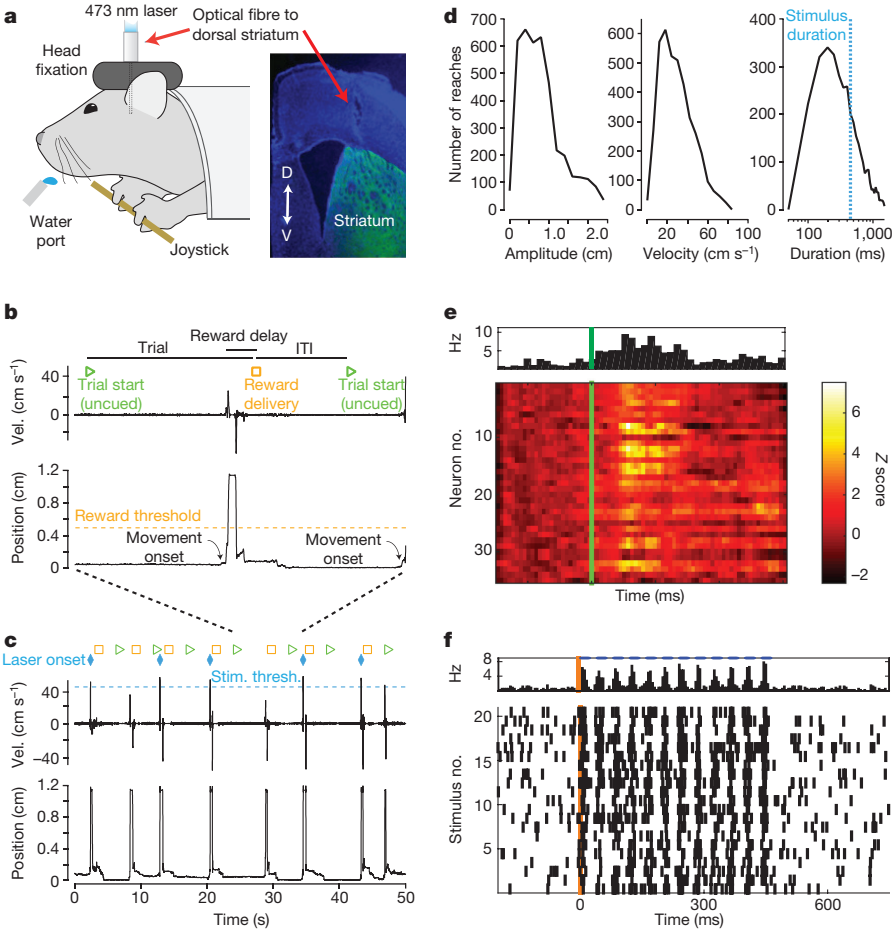


Figure 1 | Paradigm for closed-loop stimulation in dorsomedial striatum. a, Mice were head-fixed in front of a side-mounted joystick and a water port. Optical fibres were chronically implanted. Tips were positioned in the dorsomedial striatum and coupled to a 473 nm laser. Insert shows fibre position; D, dorsal; V, ventral. Fluorescent image is from iMSN neurons expressing ChR2–YFP. **b**, To receive liquid reward, mice made forelimb movements with the joystick (either a pull or push) past the criterion distance. Reward delivered 1 s after threshold crossing. Inter-trial intervals (ITIs) were 3 s (uncued). **c**, Instantaneous velocity and position of joystick for seven trials (green triangle indicates trial start). Velocity threshold for closed-loop optical stimulation and time of stimulation onset indicated by the blue dashed line and diamonds, respectively. Yellow squares indicate reward. **d**, Histograms of movement amplitude, peak velocity, and duration for all eight mice (45 sham sessions). **e**, Average response (z-scored change from baseline firing rate) of striatal units aligned to movement onset from a single session. Population average shown above. **f**, Raster plot of population activity during photostimulation from a single session.

confined to the stimulated trial[13,14] or concomitant with stimulus delivery[11].

If stimulation of the fastest movements produces a persistent change in the selection of movement parameters, the change should persist without stimulation. We plotted the velocity of movements made during the block of trials immediately following the stimulation block. In this recovery block, no stimulation was delivered. We found that stimulation-induced changes in the distribution of velocities persisted for tens of trials before gradually returning to the pre-stimulation baseline during the recovery block (Fig. 2a, d; paired t-test, $P = 0.64$ and 0.90, dMSN and iMSN, respectively). Importantly, this return to the pre-stimulation distribution had a similar time course whether it required a decrease or increase in the mean velocity following dMSN or iMSN stimulation, respectively.

We have shown that dMSN and iMSN have opponent roles in the reinforcement of movement parameters with unprecedented specificity. Stimulation induced changes are signed—dMSN stimulation increases a kinematic parameter of movement (velocity) whereas iMSN stimulation decreases the same property. However, there is a limitation to this simple opponency for learning: reinforcement should, in principle, alter behaviour so as to increase a reinforcing outcome regardless of the sign of the behavioural change[16]. It should be possible, for example, to learn to move more slowly to obtain more reward. Our data are also consistent with an alternative possibility: dMSN stimulation may be sufficient to drive changes towards movements that elicit stimulation independent of the sign (positive/negative) of the change. To distinguish between these alternatives, we stimulated MSNs during the slowest, rather than the fastest, third of limb movements. This stimulation protocol produced the opposite effects for both dMSN and iMSN stimulation (Fig. 2e, f). Under these conditions, stimulation of dMSN was sufficient to produce a cumulative decrease in velocity ($-1.1\ \mathrm{cm\ s^{-1}}$, $P = 0.008$). Conversely, iMSN stimulation produced an accumulating increase in velocity ($0.9\ \mathrm{cm\ s^{-1}}$, $P = 0.012$). Thus, the direct and indirect pathways of the basal ganglia are opponent pathways that are also sufficient for bidirectional changes in a continuous parameter that specifies purposive movement.

Models of the basal ganglia in which reinforcement learning acts to select amongst mutually exclusive actions can explain a broad array of empirical results in the learning literature[12]. However, such models cannot readily account for reinforcement acting on a continuous parameter of movement such as velocity[12] (see Supplementary Discussion). By contrast, a learning rule in which closed-loop stimulation provides a pathway-specific, signed learning signal that determines the mean of the velocity distribution could reproduce our data (Fig. 3a and Methods). Owing to the bidirectional behavioural changes observed, this learning rule makes a specific prediction: stimulation on every trial

© 2016 Macmillan Publishers Limited. All rights reserved

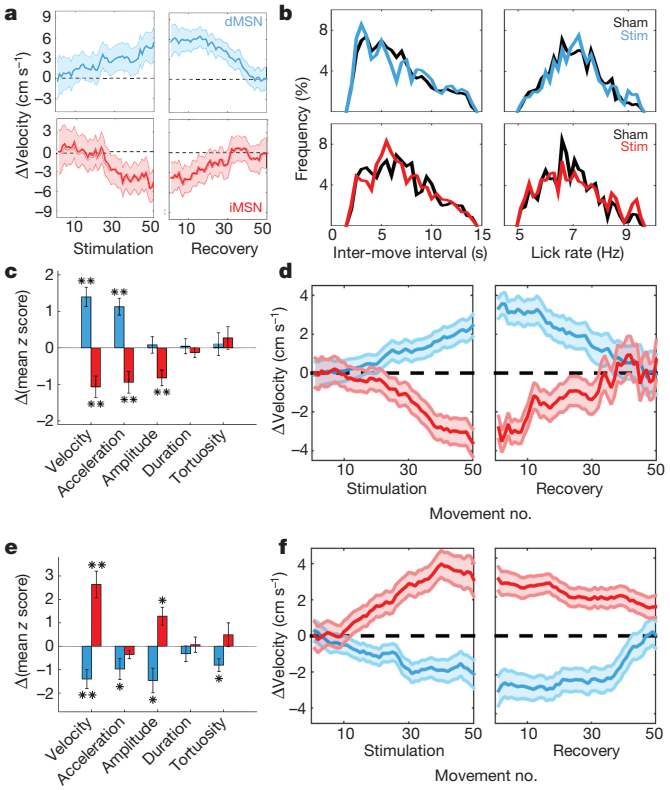


Figure 2 | Closed-loop stimulation produces opponent, bidirectional control of movement velocity. a, Difference in peak velocity between stimulation and sham session (Δvelocity) for sessions in which dMSN (upper, blue throughout) or iMSN (lower, red throughout) were stimulated on the fastest third of 50 trials during stimulation and no stimulus was delivered during recovery. Example session shown. **b**, Histograms of inter-movement-interval (left) and lick rate during reward consumption (right) for sham (black; 25 sessions in 4 dMSN mice, 20 sessions in 4 iMSN mice) and stimulation (coloured; 22 sessions in dMSN mice, 26 sessions in iMSN mice) sessions. **c**, Population average of change in movement parameters when fastest third of reaches were stimulated. **d**, Population average Δvelocity as a function of movement (trial) number when fastest third of reaches were stimulated. **e**, **f**, Same as **c**, **d**, but for sessions in which stimulation occurred on the slowest third of movements. $*P<0.05$; $**P<0.005$, two tailed t-test. Shaded area indicates standard error of the mean. Data are from 16 stimulation and 18 sham sessions in the same 4 dMSN mice, 20 stimulation, 16 sham sessions in the same 4 iMSN mice.

or at random throughout a session should produce no net change in velocity. Consistent with this prediction, each simulated stimulation protocol failed to produce a detectable change in movement velocity ($P>0.2$ for all conditions, Fig. 3 and Extended Data Fig. 5).

As formulated, this learning rule would induce a persistent change in velocity following stimulation. Extinction formulated as a fixed decay in synaptic weight[12] would not produce symmetric recovery as observed (Fig. 2 and Supplementary Discussion). To account for this feature of the data, we assumed a homeostatic component and refer to the rule as 'mean shift with homeostasis' (MeSH). Thus, the mean velocity of movement is determined by a set point that opposes learned changes and restores velocity towards baseline during recovery. When incorporated into the learning rule, we found that simulations closely reproduced the data during stimulation and recovery epochs. Selective stimulation that biased the reward-based feedback steadily drove velocity towards (dMSN) or away (iMSN) from the threshold that elicited stimulation (Fig. 3a). Upon cessation of stimulation, the homeostatic component drove recovery to the pre-stimulation baseline within 50 trials (Fig. 3).

MeSH assumes an explicit interaction between reward signalling, putatively carried by midbrain dopaminergic inputs to the dorsal striatum[6], and exogenous activation of MSNs. In contrast to this prediction, previous work has suggested that intracranial self-stimulation supported by striatal stimulation is independent of dopamine receptor activation[13]. However, the movement-related activity of striatal populations and our use of brief 450 ms stimulation both differ from the sustained, post-movement stimulation used previously. Thus, we next asked if dopamine was necessary for stimulation to elicit changes in movement velocity. We found that a low concentration of D1 and D2 receptor antagonists (SCH23390, 0.02 mg kg^{-1} and sulpiride, 25 mg kg^{-1}) injected before a behavioural session[13] eliminated persistent changes in velocity following closed-loop stimulation of either dMSN or iMSN (Fig. 3b, d) while largely sparing normal task performance (Fig. 3c, all dMSN parameters $P>0.2$; all iMSN parameters $P>0.3$). Dopamine antagonists significantly reduced the magnitude of the stimulation effect for both dMSN stimulation (92% decrease; 1.3 cm s^{-1}, $P<1\times10^{-8}$) and iMSN stimulation (109% increase; 1.2 cm s^{-1}, $P=1.5\times10^{-7}$).

Our results suggest that stimulation engages a dopamine-dependent, bidirectional plasticity in striatal MSNs. While a learning rule that acts directly on a parameter specifying the velocity distribution can account for our behavioural results, MeSH is abstract and it is unclear how it could be implemented in corticostriatal circuits critical for goal-directed, instrumental behaviour[17]. Thus, we implemented a simplified corticostriatal circuit model with the following key features (Extended Data Fig. 6).

While an action-value formulation implies that movements of different velocities are represented as a set of distinguishable neural states, MeSH implies a continuous representation of speed. There is little empirical evidence for representation of specific velocity ranges in cortical activity[18]. By contrast, there is substantial evidence that cortical[19,20] and striatal[21] representations of forelimb movements are monotonically tuned to speed. Consistent with the anatomy of the corticostriatal pathway[1], we assume that the speed of a movement is determined by both cortical and basal ganglia output (Extended Data Fig. 6). In combination with monotonic tuning this implies that the mean movement velocity is proportional to the average weight of corticostriatal synapses (Methods).

Dopamine and spike-timing dependent plasticity (STDP) has been described in the striatum[22,23]. STDP can result in bidirectional plasticity with the balance of potentiation and depression adapted to the range of population activity (that is, BCM-type plasticity[24]) in the presence of variable spike trains[25]. We assume a balance of potentiation and depression such that movements made at the average speed produce no net change. We posit that photostimulation enhances both potentiation and depression consistent with our observation that selective stimulation produces biased changes whereas non-selective stimulation produces no net change (Figs 2 and 3; Supplementary Discussion). Balanced synaptic plasticity that is enhanced by stimulation is sufficient to produce bidirectional changes in the average corticostriatal synapse weight during selective photostimulation. When incorporated into a corticostriatal circuit model, this plasticity rule produces opponent, bidirectional, and symmetric changes in movement speed (Fig. 4a, b and Extended Data Fig. 6).

Finally, we sought to validate our circuit model with electrophysiological recordings from dorsomedial striatum. Individual units (putative MSNs) recorded from mice performing the task were monotonically tuned to movement velocity (Fig. 4c), confirming previous observations[21,26]. An important feature of our behavioural results was that closed-loop stimulation does not simply reinforce stimulated movements, but rather produces a change in the mean velocity. Thus, changes in striatal activity should be apparent even for unstimulated movements. Specifically, the slope relating firing rate to velocity should be changed (Fig. 4b). Moreover, if photostimulation is necessary to alter plasticity in the recorded neuron then slope changes should correlate with photostimulation (Fig. 4b). To test this prediction we analysed a population of striatal units ($n=35$) during closed-loop stimulation of

© 2016 Macmillan Publishers Limited. All rights reserved

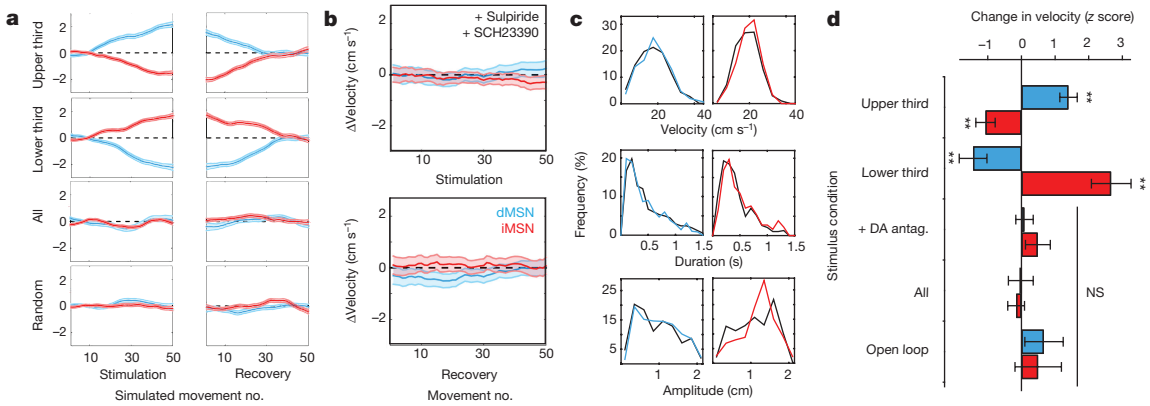


Figure 3 | Changes in velocity are consistent with dopamine-dependent reinforcement learning. a, Simulation of MeSH learning rule (see text for details). Change in average peak velocity (arbitrary units) as a function of trial number for dMSN-stimulation (blue) and iMSN-stimulation (red) simulations. **b**, ΔVelocity as a function of trial for stimulation of dMSN (blue) and iMSN (red) on the fastest third of 50 stimulation trials in the presence of dopamine receptor antagonists. Data from 14 stimulation and 11 sham dMSN sessions; 8 stimulation and 9 iMSN sham sessions. **c**, Movement parameter distributions for control sessions (black) and sessions following dopamine antagonist administration (coloured). **d**, Summary of the changes in velocity for experiments as indicated for dMSN (blue) and iMSN (red) stimulation sessions as defined in text. DA antag., dopamine receptor antagonist. Shaded area and error bars indicate standard error of the mean. $^{**}P < 0.005$, two tailed *t*-test; NS, not significant.

dMSNs on the fastest third of movements. Consistent with the model predictions, we observed an increase in the slope of the velocity tuning associated with an increase in the average velocity (Fig. 4d). Changes in tuning tended to be most prominent on units responsive to photostimulation (putative dMSNs; Fig. 4d) whereas neurons with weak stimulation responses (putative iMSNs or distant dMSNs) tended to show decreases in apparent tuning slope (Fig. 4d).

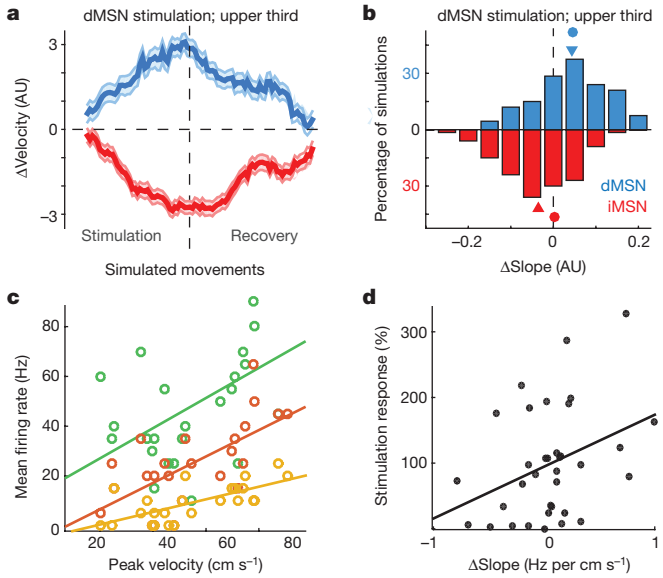


Figure 4 | Corticostriatal circuit model implements MeSH rule and experimental validation. a, Change in movement velocity (arbitrary units, AU) as a function of number of simulated trials (shaded area indicates standard deviation; $n = 100$) for simulations in which dMSN (blue) or iMSN (red) were stimulated during fastest ~30% of reaches. **b**, Histogram of the change in the average slope between dMSN (blue) or iMSN (red) activity and movement velocity for simulations with dMSN stimulation. Triangles indicate mean. Circles indicate mean change in weight of corticostriatal synapse (see text for details). **c**, Mean firing rate during movement (0–505 ms after onset) is plotted as a function of peak movement velocity with overlaid linear fit ('tuning slope') for 3 of 35 example striatal units. **d**, Stimulation response (per cent change in baseline firing rate during photostimulation) as a function of the change in the tuning slope (Δslope) for unstimulated trials (subthreshold velocity) of sessions in which dMSN were stimulated on the fastest third of movements. Linear fit overlaid ($r = 0.47$).

Here we provide the first demonstration, to our knowledge, that the innate biases in the direct and indirect pathway to increase or decrease the frequency of movement, respectively, do not extend to fixed biases in the control of movement parameters. The direct and indirect pathways engage opponent, activity-dependent plasticity mechanisms that can produce sustained biases in future behaviour. Each pathway is sufficient to produce bidirectional changes and, to some extent, is innervated by distinct cortical populations[27], suggesting that bidirectional control by each pathway could allow for adaptive control of goal-directed actions in different contexts or under different demands. These data argue that phasic activity in the striatum during specific movements is sufficient to selectively reinforce changes in a movement parameter independent of a generalized change in motivation consistent with a role for dopamine-dependent signalling in dorsal striatum in the control of movement vigour[21,26,28].

Our results reveal a bidirectional control of behaviour by MSNs that may be contrasted with the recent observation that self-stimulation supported by MSNs is unidirectional and dopamine-independent[13]. The differences between the findings may reflect the different experimental paradigms. Selectively biasing striatal activity in the context of a reward-based operant task could engage mechanisms distinct from the reinforcing properties of photostimulation itself. In the latter case strong stimulation may be sufficient to replace dopaminergic inputs or support self-stimulation in a dopamine-independent manner[29]. In addition, we observed a symmetric recovery to baseline following cessation of stimulation that is also distinct from the differential extinction of self-stimulation[13]. However, a recent modelling study argued that apparent differences in extinction are consistent with equivalent learning rates following dMSN and iMSN stimulation[12] consistent with our observation of opponent, but symmetric effects.

Here we proposed a circuit implementation by which a continuous parameter defining a purposive movement can be selectively reinforced by a stimulation-dependent enhancement of bidirectional synaptic plasticity. Importantly, it has been shown that striatal neurons are capable of bidirectional synaptic plasticity[22,23]; however, plasticity is mediated by distinct signalling events in the two populations[23]. Resolving the roles played by the intersection of these different cellular and circuit factors that govern bidirectional plasticity will be critical to understand the role of dopamine in instrumental learning. In addition to kinematic parameters of movement, other aspects of reinforcement learning

© 2016 Macmillan Publishers Limited. All rights reserved

are governed by continuous parameters such as rates[30] or value[6]. The circuit implementation we propose could provide a general mechanism by which activity-dependent plasticity in striatum produces learned changes in continuous parameters with monotonic representations in neural activity.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 15 May 2015; accepted 3 March 2016.
Published online 2 May 2016.

1. Dudman, J. T. & Gerfen, C. R. in *The Rat Nervous System* (ed. G Paxinos) Ch. 17, (Elsevier, 2015).
2. Balleine, B. W., Liljeholm, M. & Ostlund, S. B. The integrative function of the basal ganglia in instrumental conditioning. *Behav. Brain Res.* **199,** 43–52 (2009).
3. Mink, J. W. The basal ganglia: focused selection and inhibition of competing motor programs. *Prog. Neurobiol.* **50,** 381–425 (1996).
4. Frank, M. J. Computational models of motivated action selection in corticostriatal circuits. *Curr. Opin. Neurobiol.* **21,** 381–386 (2011).
5. Gurney, K. N., Humphries, M. D. & Redgrave, P. A new framework for cortico-striatal plasticity: behavioural theory meets *in vitro* data at the reinforcement-action interface. *PLoS Biol.* **13,** e1002034 (2015).
6. Schultz, W. Behavioral theories and the neurophysiology of reward. *Annu. Rev. Psychol.* **57,** 87–115 (2006).
7. Desmurget, M. & Turner, R. S. Motor sequences and the basal ganglia: kinematics, not habits. *J. Neurosci.* **30,** 7685–7690 (2010).
8. Tumer, E. C. & Brainard, M. S. Performance variability enables adaptive plasticity of 'crystallized' adult birdsong. *Nature* **450,** 1240–1244 (2007).
9. Andalman, A. S. & Fee, M. S. A basal ganglia-forebrain circuit in the songbird biases motor output to avoid vocal errors. *Proc. Natl Acad. Sci. USA* **106,** 12518–12523 (2009).
10. Gerfen, C. R. *et al.* D1 and D2 dopamine receptor-regulated gene expression of striatonigral and striatopallidal neurons. *Science* **250,** 1429–1432 (1990).
11. Kravitz, A. V. *et al.* Regulation of parkinsonian motor behaviours by optogenetic control of basal ganglia circuitry. *Nature* **466,** 622–626 (2010).
12. Collins, A. G. & Frank, M. J. Opponent actor learning (OpAL): modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychol. Rev.* **121,** 337–366 (2014).
13. Kravitz, A. V., Tye, L. D. & Kreitzer, A. C. Distinct roles for direct and indirect pathway striatal neurons in reinforcement. *Nature Neurosci.* **15,** 816–818 (2012).
14. Tai, L. H., Lee, A. M., Benavidez, N., Bonci, A. & Wilbrecht, L. Transient stimulation of distinct subpopulations of striatal neurons mimics changes in action value. *Nature Neurosci.* **15,** 1281–1289 (2012).
15. Niv, Y., Daw, N. D., Joel, D. & Dayan, P. Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology (Berl.)* **191,** 507–520 (2007).
16. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: an Introduction.* (MIT Press, 1998).
17. Yin, H. H., Ostlund, S. B., Knowlton, B. J. & Balleine, B. W. in *Eur. J. Neurosci.* **22,** 513–523 (2005).
18. Paninski, L., Fellows, M. R., Hatsopoulos, N. G. & Donoghue, J. P. Spatiotemporal tuning of motor cortical neurons for hand position and velocity. *J. Neurophysiol.* **91,** 515–532 (2004).
19. Churchland, M. M. & Shenoy, K. V. Temporal complexity and heterogeneity of single-neuron activity in premotor and motor cortex. *J. Neurophysiol.* **97,** 4235–4257 (2007).
20. Moran, D. W. & Schwartz, A. B. Motor cortical representation of speed and direction during reaching. *J. Neurophysiol.* **82,** 2676–2692 (1999).
21. Panigrahi, B. *et al.* Dopamine is required for the neural representation and control of movement vigor. *Cell* **162,** 1418–1430 (2015).
22. Pawlak, V. & Kerr, J. N. Dopamine receptor activation is required for corticostriatal spike-timing-dependent plasticity. *J. Neurosci.* **28,** 2435–2446 (2008).
23. Shen, W., Flajolet, M., Greengard, P. & Surmeier, D. J. Dichotomous dopaminergic control of striatal synaptic plasticity. *Science* **321,** 848–851 (2008).
24. Cooper, L. N. & Bear, M. F. The BCM theory of synapse modification at 30: interaction of theory with experiment. *Nature Rev. Neurosci.* **13,** 798–810 (2012).
25. Izhikevich, E. M. & Desai, N. S. Relating STDP to BCM. *Neural Comput.* **15,** 1511–1523 (2003).
26. Turner, R. S. & Desmurget, M. Basal ganglia contributions to motor control: a vigorous tutor. *Curr. Opin. Neurobiol.* **20,** 704–716 (2010).
27. Wall, N. R., De La Parra, M., Callaway, E. M. & Kreitzer, A. C. Differential innervation of direct- and indirect-pathway striatal projection neurons. *Neuron* **79,** 347–360 (2013).
28. Mazzoni, P., Hristova, A. & Krakauer, J. W. Why don't we move faster? Parkinson's disease, movement vigor, and implicit motivation. *J. Neurosci.* **27,** 7105–7116 (2007).
29. Phillips, A. G. & Fibiger, H. C. The role of dopamine in maintaining intracranial self-stimulation in the ventral tegmentum, nucleus accumbens, and medial prefrontal cortex. *Can. J. Psychol.* **32,** 58–66 (1978).
30. Gallistel, C. R. & Gibbon, J. Time, rate, and conditioning. *Psychol. Rev.* **107,** 289–344 (2000).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by funding from the Howard Hughes Medical Institute. J.T.D. is a Group Leader at Janelia Research Campus. We thank A. Lee, A. Karpova, N. Spruston, and members of our laboratory for critical reading and feedback on the manuscript. We also thank M. Frank for helpful discussions of the OpAL model.

Author Contributions E.A.Y. performed the experiments and analysed the data. E.A.Y. and J.T.D. designed the experiments, performed the modelling, and wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.T.D. (dudmanj@janelia.hhmi.org).

© 2016 Macmillan Publishers Limited. All rights reserved

METHODS

Subjects. Experimental subjects were eight adult (over 2 months old) male mice, four each of Drd1a-Cre (http://www.informatics.jax.org/allele/MGI:3836631) or Drd2-Cre (http://www.informatics.jax.org/allele/MGI:3836635) crossed with a mouse with an allele for Cre-dependent expression of channelrhodopsin-2 fused to enhanced yellow fluorescent protein (Ai32; https://www.jax.org/strain/012569). Mouse lines expressing Cre-recombinase were produced by the GENSAT project (GENSAT project, Rockefeller University, New York, USA)[31,32] and obtained from the MMRC (https://www.mmrrc.org). Ai32 mice were obtained from Jackson Laboratory and produced by the Allen Institute for Brain Science (https://www.alleninstitute.org)[33]. Number of animals and sessions were based upon previous studies using an intersession control model. Experimenters were not blinded to the condition or animal strain. All animals were handled in accordance with guidelines approved by the Institutional Animal Care and Use Committee (IACUC) of Janelia Research Campus which is IAAALAC accredited.

Animal care. Mice were individually housed in a temperature- and humidity-controlled room maintained on a reversed 12-h light/dark cycle. Following 1 week of recovery from surgery, the water consumption of the mice was limited to at least 1 ml per day. Mice underwent daily health checks, and water restriction was eased if mice fell below 70% of their body weight at the beginning of deprivation. Mice were acclimated to head fixation and trained to lick drops of water sweetened with saccharin.

Behavioural training. Mice spent 4–8 weeks adjusting to being head-fixed and learning to displace a side-mounted joystick, placed 2.7 cm away from their platform, to a threshold of roughly 0.5 cm. After this initial training, the threshold for a successful trial ('criterion threshold') was reduced 0.1 cm so that every joystick movement would be rewarded. Reaching movements were self-paced. Most movement amplitudes easily exceeded this amplitude threshold (Fig. 1c).

Mice were restricted to consume 1.5 ml of water per day to maintain motivation for task completion. Movement was measured by recording voltage changes applied across the variable potentiometer connected to the joystick and were found to be linearly proportional to displacement over the range of movement amplitudes used by mice. At the start of each trial, joystick position was centred to coordinates (0,0), and animals were trained to manoeuvre the joystick to certain displacement thresholds equal to a set resistance change of a potentiometer. Both movements away from and towards the body were allowed. Delivery of a sweetened water reward (~0.05–0.1 ml per trial; controlled by an audible solenoid valve) signalled a successful movement and advancement to the next trial. Water delivery was delayed by 1,000 ms after the joystick position crossed a specific distance threshold. The threshold was set at an arbitrary, low value such that false positives were not detected, but all trial-initiating movements in well-trained mice were rewarded. A force of 0.1 N was required to displace and hold the joystick at an eccentric position. For reference, this is at least fivefold less than a mouse can pull towards itself using its forelimbs for several seconds[34].

No other task-related stimulus was present and the behaviour was performed in a darkened behaviour box. Rewards were followed by a 4,000 ms inter-trial interval (ITI) in which no movements would be rewarded. The joystick position at the end of the ITI (almost always near the central default position) was used as the initial position for the subsequent trial. Mice performed at least 125 trials per session. The initial 25 trials were only used for daily acclimation of the animal to the behavioural setup. Blocks of 50 trials were performed with the stimulation block followed by the no stimulation block. Many blocks could be completed, but only the first block from each condition were used in these analyses. Sham stimulation sessions were identical to stimulation sessions, including the attachment of the optic fibre to the mouse's head, with the exception that the laser was not turned on. Stimulation sessions were conducted at random, and the data from the subsequent session were not used in analysis.

Fibre implantation and optical stimulation. Implantation surgery was performed under full anaesthesia (1.5% isoflurane). The skull was exposed and fibre optic probes were unilaterally inserted 2.2 mm into the brain at 0.5 mm anterior and 1.8 mm lateral to bregma. Fibre optic probes were made of glass fibres (100 μm core) fitted with zirconia LC connectors. Head fixation caps were implanted at the end of the procedure and all elements and remaining skull were covered with dental acrylic as described previously[35]. All surgical procedures were performed under aseptic conditions.

Fibre implants were targeted to the dorsomedial aspect of the striatum (DMS)[1]. Extended Data Fig. 1a shows the localization of the tips of optical fibres implanted for dMSN and iMSN stimulation. To characterize this specific location in more detail we performed bilateral injections of a retrograde tracer (Lumiflor beads) into the approximate DMS location of fibre implants. We found extensive retrograde labelling of neocortical neurons over a relatively extended rostro-caudal axis that was biased towards the medial aspect of neocortex (Extended Data Fig. 1b, c), consistent with previous results from our lab[36]. Based upon the anatomical atlas of the mouse brain[37] these cortical structures are annotated as M2 (secondary motor cortex) and Cg (cingulate cortex). However, we note that recent functional mapping of neocortex indicates that these sites are also within the boundaries of the rostral and caudal forelimb regions (Extended Data Fig. 1c)—areas that are sufficient to produce forelimb movements in response to microstimulation[38].

Closed-loop photostimulation on high- and low-velocity blocks was accomplished through online monitoring of instantaneous joystick velocity. Thresholds for triggering the laser were set for each animal such that approximately one-third of baseline movements would be suprathreshold. The velocity threshold within a session was fixed, but on occasion it was changed from one session to the next. Thresholds for all mice and all sessions were within 6% of each other. For stimulation of all movements, we set the velocity threshold low enough that all movements were suprathreshold. For low-velocity triggering (Fig. 3), we took advantage of the reliable, stereotyped nature of the reaches to predict peak velocity from early velocity. To trigger photostimulation, velocity needed to initially pass a low, 'onset' threshold while not exceeding a higher 'too fast' threshold for the next 20 ms. Using our real-time velocity triggering, we correctly stimulated 96% of upper-third (fast) reaches protocol and 84% of lower-third (slow) reaches protocol. Our false positive stimulation rate was 9% for both protocols. Photostimulation consisted 10-ms pulses at 16.7 Hz for 450 ms from a 473 nm blue laser set so that the power at the tip of the implanted optic probe was 3–6 mW. This was at a frequency below that which individual neurons could reliably follow (Extended Data Fig. 7). Upper-third stimulation data consist of 22 stimulation and 25 sham sessions in dMSN mice, 26 stimulation and 20 sham sessions in iMSN mice. Lower-third stimulation data consist of 16 stimulation and 18 sham sessions in dMSN mice, 20 stimulation, 16 sham sessions in iMSN mice.

We next sought to estimate the extent of light spread based upon the laser power, fibre diameter and duty cycle of our pulse train using a combination of simulation[39] and electrophysiology. The simulation result is shown in Extended Data Fig. 1d. In brief, the peak intensity of light stimulation was reduced to 1% maximum by ~1 mm below and 0.5 mm lateral to the optical fibre. To directly estimate the change in stimulation efficacy as a function of distance we performed recordings with a 4-shank silicon probe (NeuroNexusTech; 'Buzsaki32' site layout) on which 1 shank was affixed with an optical fibre. Consistent with the estimate of light scattering from the simulation we found that direct light activation was substantially weaker on the neighbouring shanks (Extended Data Fig. 1e). At the location of our fibres (Extended Data Fig. 1) the dorsal striatum extends for ~2 mm is all dimensions and the DMS roughly extends for 1 mm. Thus, these data indicate that direct photostimulation was restricted to the dorsal striatum.

Behavioural analysis. All behavioural events were recorded on separate channels at 1 kHz (BlackRock Microsystems; Salt Lake City, Utah). Data analysis was performed using written routines in MATLAB 2014 (MathWorks; Natick, Massachusetts) to extract individual forelimb movement trajectories (Extended Data Fig. 8). Quantification of individual movements considered only the outward component of the reach and quantified the peak amplitude and velocity. The beginning of the reach was assessed offline for each reach and was determined to be the first time point constituting the increasing velocity associated with that reach. The duration was computed as the full duration of the movement and tortuosity is a measure of the directness of the reach path, defined as the path length divided by the end point distance. Z scores were computed for each stimulation session, using the average sham session mean and standard deviation, within each animal, then combined across animals. Simulations of behavioural learning were implemented in MATLAB and are described in detail in the Extended Data. Unless otherwise noted, statistical significance refers to $P < 0.05$, two-tailed Student's t-test.

Electrophysiology. Extracellular electrophysiology was performed in the dorsal striatum of awake, behaving mice. 32 channel silicon probe arrays with attached integrated optical fibres (NeuroNexus Tech; 'Buzsaki32' site arrangement) were acutely implanted in the dorsal striatum (centre of array was positioned 0.5 mm anterior and 1.8 mm lateral to bregma and −2.0 mm to −3.0 mm depth from surface). Electrodes were prepared for recording by reducing the site impedance below 750 kOhm. Broadband continuous data (0.1 Hz–7.5 kHz) were recorded with simultaneous sampling of voltage from the joystick, the lick port, and digital signals from the behaviour control system (30 kHz sample rate on all channels, Blackrock Microsystems, Salt Lake City, Utah). Continuous voltage signals were highpass filtered (0.5–7 kHz) offline and events that exceeded four times the standard deviation of the continuous voltage signal were extracted (spikes). Spike sorting into individual units was performed in MATLAB using custom-written software. Spikes were isolated according to waveform amplitude distribution and principal components of the amplitude array across the eight electrodes (~25 μm spacing) of each shank ($n = 8$) of the silicon probe array. The event times for each individual single unit were then aligned to movement start as extracted from the continuous

© 2016 Macmillan Publishers Limited. All rights reserved

voltage signal from the joystick. Velocity–firing-rate slopes were computed using the mean activity of each unit over the epoch spanning 0 to 400 ms after reach initiation. The evoked response for each stimulus (Fig. 1f) was averaged across all isolated neurons, with spikes placed into 2-ms-wide bins.

Pharmacology. We injected D1 and D2 receptor antagonists (SCH23390, 0.02 mg kg^{-1} and sulpiride, 25 mg kg^{-1}, co-injected intraperitoneally) before stimulation sessions. Sham sessions were ones in which the same animals received 0.9% saline injection instead of drug.

The MeSH learning rule. To determine whether the changes in reach velocity due to stimulation were consistent with a reinforcement learning rule, we developed a simple computational model:

$$M_{i+1} = M_i + \omega_r\,(m_i - \overline{M_i}) + \omega_S S - \omega_p\,(m_i - P)/|m_i - P|$$

where M is the Gaussian distribution of values m for a given movement parameter, from which m_i is chosen at random on trial i. Performance reinforcement in the form of reward, r, shifts the mean of M relative the reward r, to which is always given and therefore will be present in every trial. Additional stimulation-induced reinforcement also occurs, shifting the mean according to the type of stimulation, S (+1 or −1 for dMSN or iMSN, respectively; 0 for no stimulation) at a fixed proportion of the reward rate, that is, $\omega_S = C_s\omega_r$. Finally, we have added a restorative set point, P, which is based upon the original mean of the distribution.

Corticostriatal circuit model. We implemented a simple simulation of 500 D1 and 500 D2 striatal units. Activity was continuously varied between 0 and 1. The activity of a given unit was defined as: $r_i = w_i \times$ active, where active $\varepsilon[0,1]$.

Movement velocity was a product of the total cortical output summed with the net contribution of striatal activity (see schematic in Extended Data Fig. 6). This is an explicit model of the structure of the corticostriatal projection where striatal neurons receive collateral input from corticocortical and coorticofugal outputs from neocortex[1]. Thus, movement velocity was: mvmt_velocity $= \alpha\sum$active $+$ $\beta\sum$d1_msn $-\gamma\sum$d2_msn. Simulations were conducted with a variety of weightings, but for examples in the manuscript we used $\alpha = 0.5$, $\beta = 1$, $\gamma = 1$.

Synapse weights were updated incrementally according to a simple update equation: $w_i(t+1) = w_i(t) +$ active $\times\,\alpha_{\text{learn}} - (1 -$ active$)\beta_{\text{learn}}$.

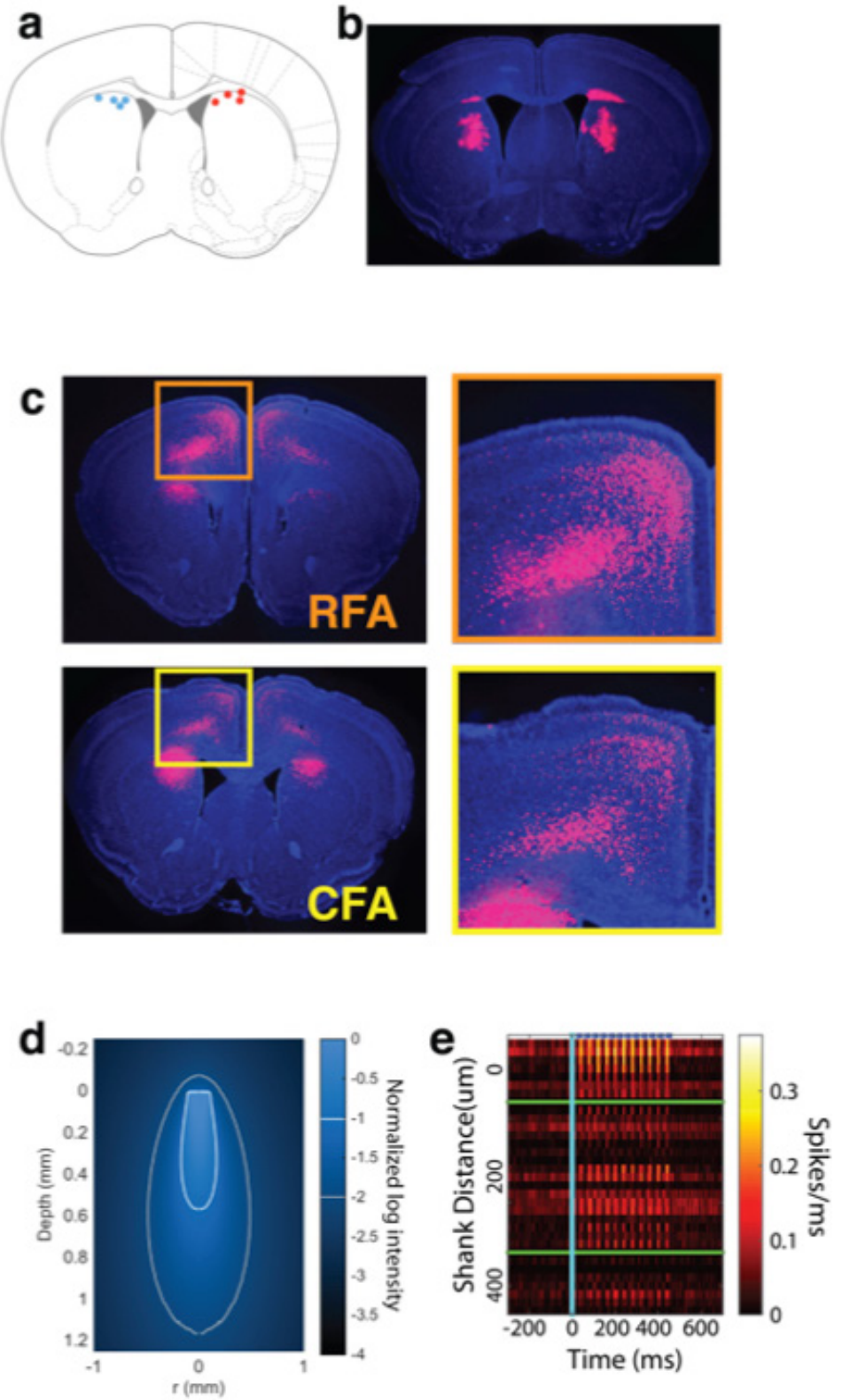
Thus, if the unit was active its weight was increased by α_{learn} or otherwise decreased by β_{learn} if inactive. Various parameterizations of learning rates could be used, but we typically used $\alpha_{\text{learn}} = -\beta_{\text{learn}} = 0.05$ in the absence of photostimulation and $\alpha_{\text{learn}} = -\beta_{\text{learn}} = 0.09$ during photostimulation. Altered learning rates were only applied to the stimulated population.

***In vitro* intracellular recordings.** Methods for Extended Data Fig. 7 were as described previously[40]. Briefly, for the preparation of *in vitro* brain slices, mice were deeply anaesthetized with isoflurane, decapitated, and the brain placed into ice-cold modified artificial cerebral spinal fluid (aCSF) (in mM: 52.5 NaCl, 100 sucrose, 26 $NaHCO_3$, 25 glucose, 2.5 KCl, 1.25 NaH_2PO_4, 1 $CaCl_2$, 5 $MgCl_2$ and in μM: 100 kynurenic acid) that had been saturated with 95%O_2/5%CO_2. 300-μm thick coronal slices were cut (Leica VT1200S; Leica Microsystems, Germany), transferred to a holding chamber and incubated at 35 °C for 30 min in modified aCSF (in mM: 119 NaCl, 25 $NaHCO_3$, 28 glucose, 2.5 KCl, 1.25 NaH_2PO_4, 1.4 $CaCl_2$, 1 $MgCl_2$, 3 Na pyruvate and in μM: 400 ascorbate and 100 kynurenic acid, saturated with 95%O_2/5%CO_2) and then stored at room temperature.

For recordings, slices were transferred to a recordings chamber and superfused with modified aCSF (in mM: 119 NaCl, 25 $NaHCO_3$, 18 glucose, 2.5 KCl, 1.25 NaH_2PO_4, 1.4 $CaCl_2$, 1 $MgCl_2$, 3 Na pyruvate and in μM: 400 ascorbate and saturated with 95%O_2/5%CO_2) maintained at 32–34 °C, at a flow rate of 2–3 ml per minute. Patch pipettes (resistance 5–8 MΩ) were pulled on a laser micropipette puller (Model P-2000, Sutter Instrument Co., Sunnyvale, California) and filled with a KGluconate-based intracellular solution (in mM: 137.5 KGluconate, 2.5 KCl, 10 HEPES, 4 NaCl, 3 GTP, 40 ATP, 10 phosphocreatine, pH 7.5). Intracellular recordings were made using a MultiClamp700B amplifier (Molecular Devices, Sunnyvale, California) interfaced to a computer using an analogue to digital converter (PCI-6259; National Instruments, Austin, Texas) controlled by custom written scripts in Igor Pro (Wavemetrics, Eugene, Oregon). Software is available at http://www.dudmanlab.org.

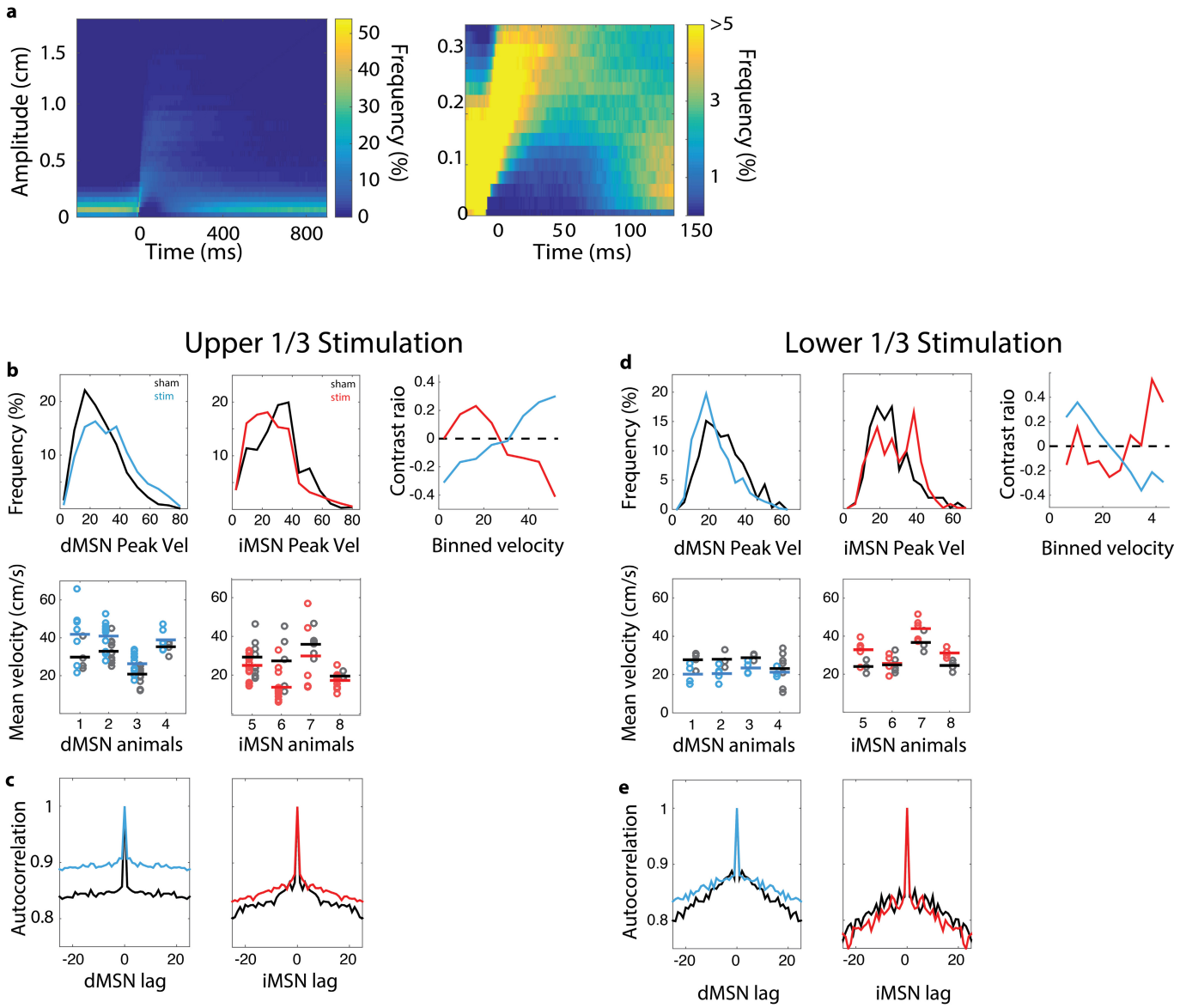
31. Gerfen, C. R., Paletzki, R. & Heintz, N. GENSAT BAC Cre-recombinase driver lines to study the functional organization of cerebral cortical and basal ganglia circuits. *Neuron* **80,** 1368–1383 (2013).
32. Gong, S. *et al.* Targeting Cre recombinase to specific neuron populations with bacterial artificial chromosome constructs. *J. Neurosci.* **27,** 9817–9823 (2007).
33. Madisen, L. *et al.* A toolbox of Cre-dependent optogenetic transgenic mice for light-induced activation and silencing. *Nature Neurosci.* **15,** 793–802 (2012).
34. Deacon, R. M. Measuring the strength of mice. *J. Vis. Exp.* (2013).
35. Osborne, J. E. & Dudman, J. T. RIVETS: a mechanical system for *in vivo* and in vitro electrophysiology and imaging. *PLoS ONE* **9,** e89007 (2014).
36. Pan, W. X., Mao, T. & Dudman, J. T. Inputs to the dorsal striatum of the mouse reflect the parallel circuit architecture of the forebrain. *Front. Neuroanat.* **4,** 147 (2010).
37. Paxinos, G. & Franklin, K. *The Mouse Brain in Stereotaxic Coordinates.* (Gulf Professional Publishing, 2004).
38. Tennant, K. A. *et al.* The organization of the forelimb representation of the C57BL/6 mouse motor cortex as defined by intracortical microstimulation and cytoarchitecture. *Cereb. Cortex* **21,** 865–876 (2011).
39. Stujenske, J. M., Spellman, T. & Gordon, J. A. Modeling the spatiotemporal dynamics of light and heat propagation for *in vivo* optogenetics. *Cell Rep.* **12,** 525–534 (2015).
40. Brown, J., Pan, W. X. & Dudman, J. T. The inhibitory microcircuit of the substantia nigra provides feedback gain control of the basal ganglia output. *Elife* **3,** e02397 (2014).

© 2016 Macmillan Publishers Limited. All rights reserved



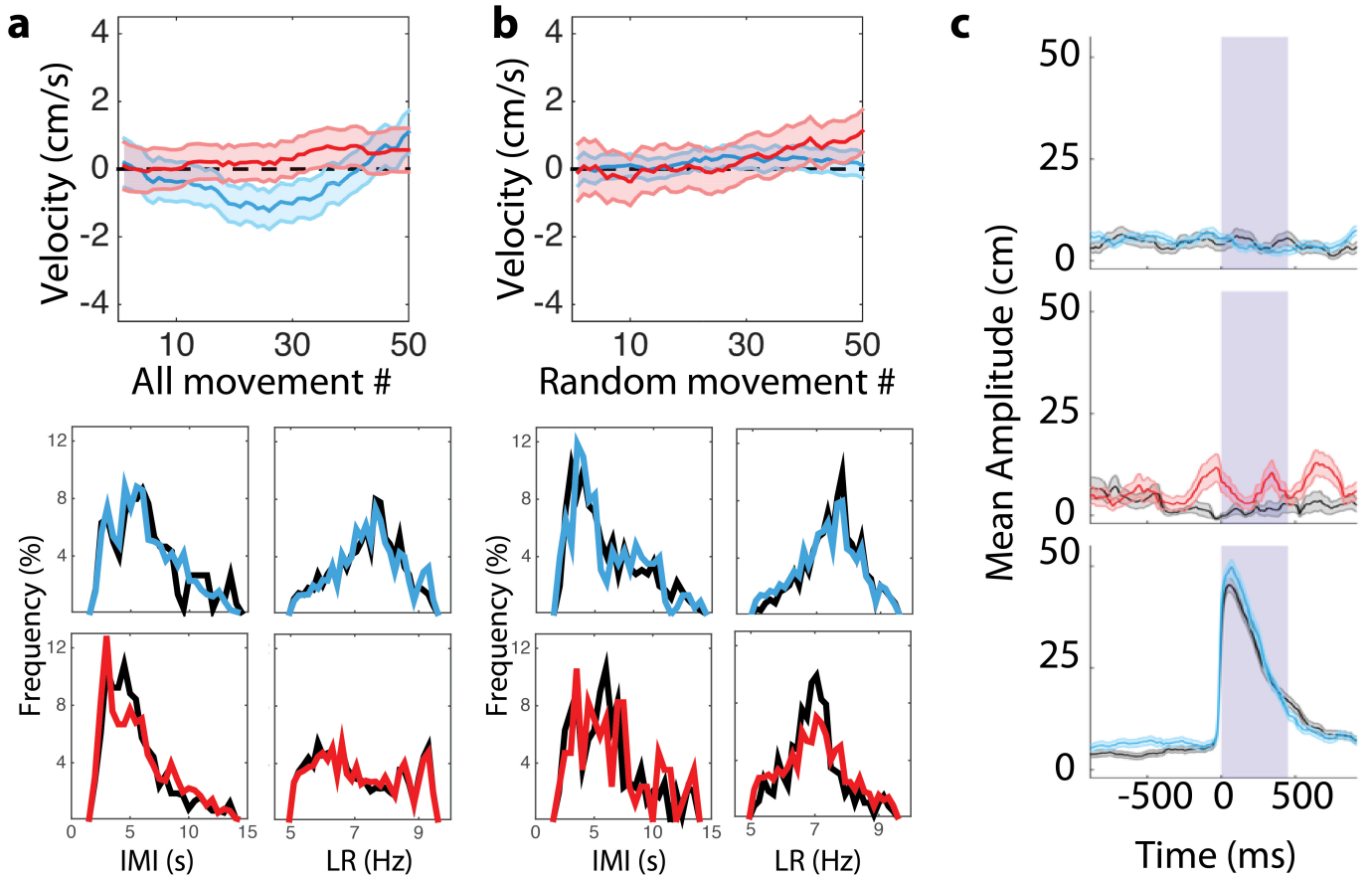
Extended Data Figure 1 | Anatomical localization of stimulated neurons and their corticostriatal inputs. a, Schematic showing the histological reconstruction of optic fibre endpoints in dMSN (cyan) and iMSN (red) mice. Hemispheric segregation for display purposes only; actual implants were randomized left and right. **b**, Location of retrobead tracer injection just below the coordinates of fibre implant terminus. **c**, Location of the labelled projection neurons innervating the retrobead injection site corresponding to the rostral forelimb area (RFA, orange) and caudal forelimb area (CFA, yellow). **d**, Results of simulation of light intensity and spread through the brain based upon the model of Stujenske *et al.*[39]. Light intensity drops below 10% peak intensity 285 μm below the fibre. At this depth the lateral fall-off is a drop to 1% peak intensity 250 μm. Contour lines for 10% intensity and 1% intensity are overlaid in white and grey, respectively. **e**, Average photostimulation-evoked activity of individual neurons (each row = 1 neuron) sorted by electrode shank. We observed minimal fidelity in the light-evoked responses of neurons on the farthest shank consistent with the predicted fall-off in light intensity from simulations. Blue dashes at top identify timing of laser pulses. Robust entrainment of spiking to photostimulation was also observed *in vitro* (Extended Data Fig. 7).

© 2016 Macmillan Publishers Limited. All rights reserved



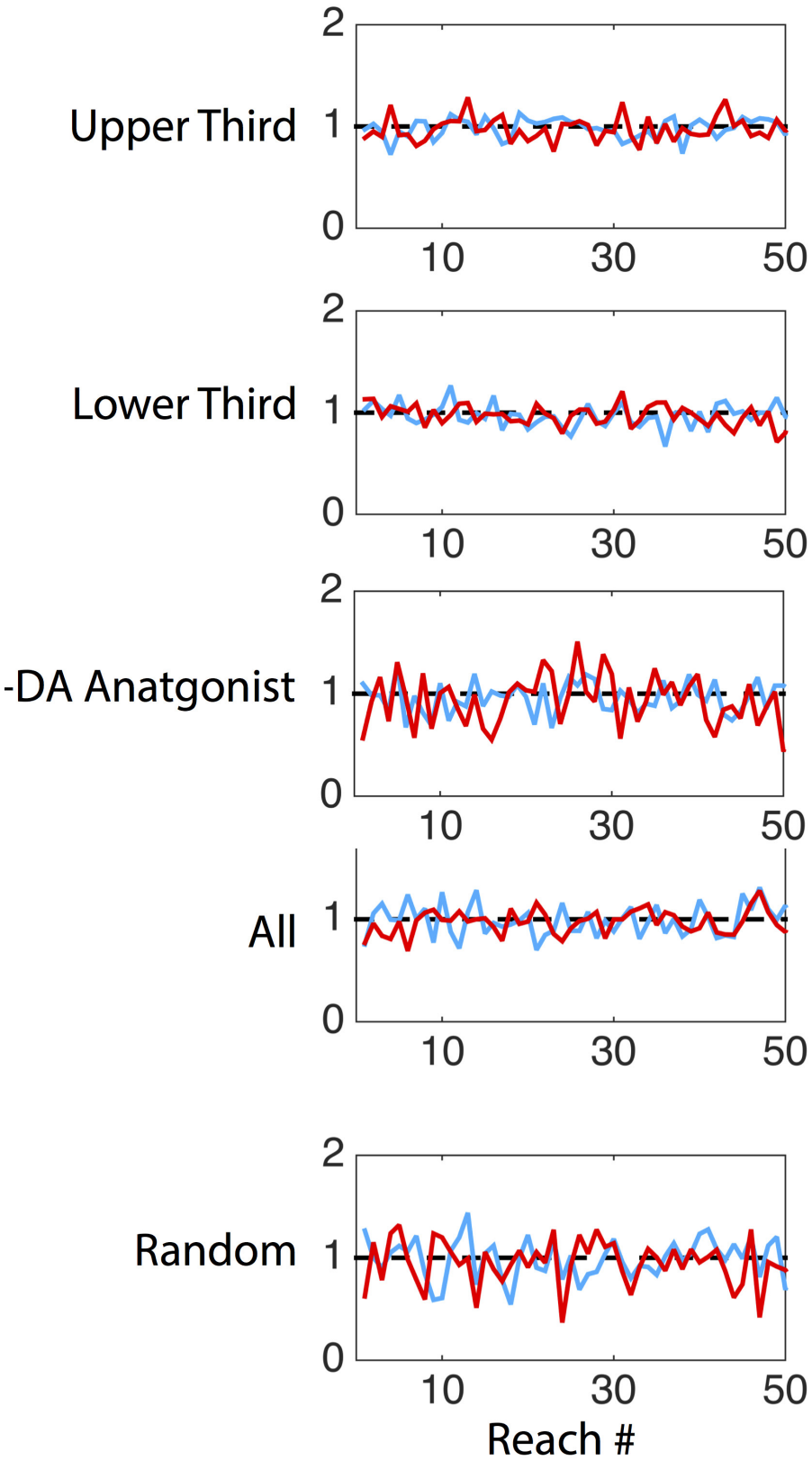
Extended Data Figure 2 | Selective stimulation of MSNs produces changes in peak velocity. a, Density plot of movement trajectories showing the percentage of movements that passed through a given amplitude throughout time, aligned to movement onset. Particularly in the magnified plot on the right, we can see that very few movement trajectories passed near the central 0.2 cm within the first 100 ms. **b**, Peak velocity distributions for sham and stimulation data sets for dMSN stimulation (blue, left column), and iMSN stimulation (red, middle column). Top row shows complete distributions across animals. Note that no particular part of the distribution is pronounced following stimulation. Bottom row shows means for all experimental sessions by animal. Animal number is for indexing purposes only. Most experiments were carried out concurrently across animals. The right column shows the contrast ratio (difference divided by sum) for dMSN and iMSN stimulation effects (blue and red, respectively). These data show a steady mean shift in the data; for instance, iMSN stimulation (red) is positive for velocities slower than the sham (control) mean (an increase in frequency), and is negative for velocity values greater than the sham mean (a decrease in frequency). We have curtailed the contrast ratio plot where too few values existed to achieve reliable estimates (55 cm s^{-1}). **c**, Autocorrelation of movement velocities for dMSN stim (blue), iMSN (red), and sham (black) data. **d**, **e**, Same analyses as **b**, **c**, but for those sessions in which stimulation occurred on the slower, lower third of movement velocities.

© 2016 Macmillan Publishers Limited. All rights reserved



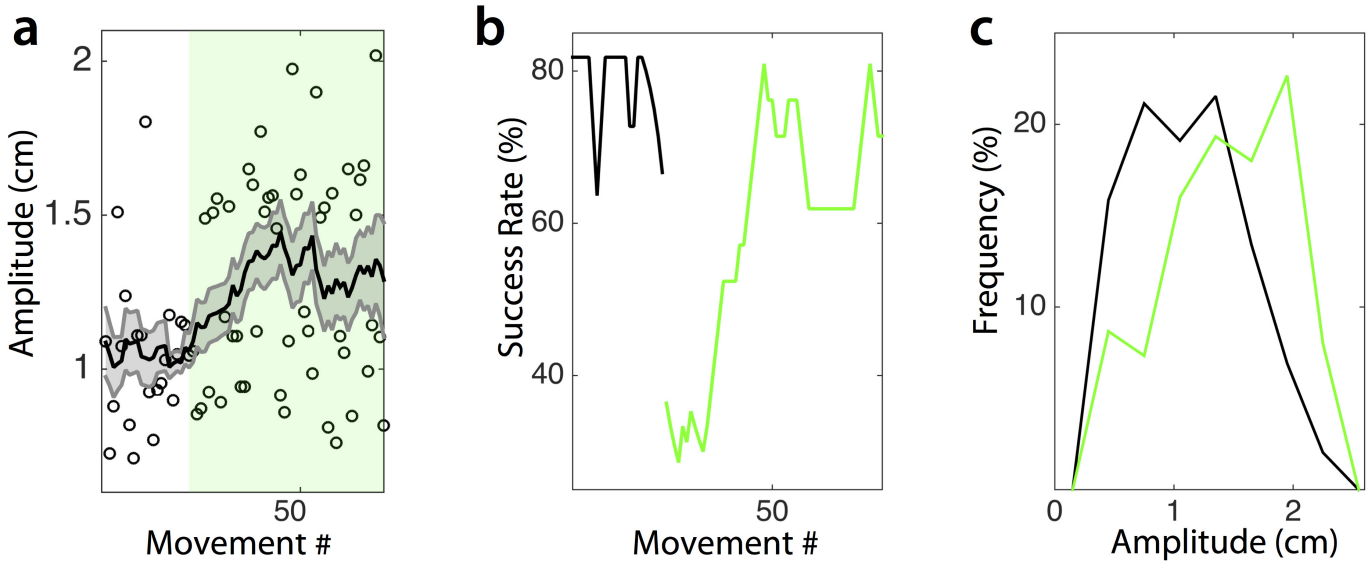
Extended Data Figure 3 | Trained animals can adjust amplitude to changing task requirements. a, Reach amplitudes from a sample of seven sessions in two mice in which the eccentricity of the threshold to receive reward was suddenly increased at random. Green field identifies reaches performed with the increased amplitude threshold. Shaded area represents standard error of the mean. **b**, Success rate before (black) and after (green) the jump in amplitude threshold. **c**, Distribution of reach amplitudes across sessions for pre- (green) and post- (black) amplitude threshold jump. These data indicate that the mean amplitude was not saturated and suggest that behaviour remains outcome-dependent (that is, goal-directed).

© 2016 Macmillan Publishers Limited. All rights reserved



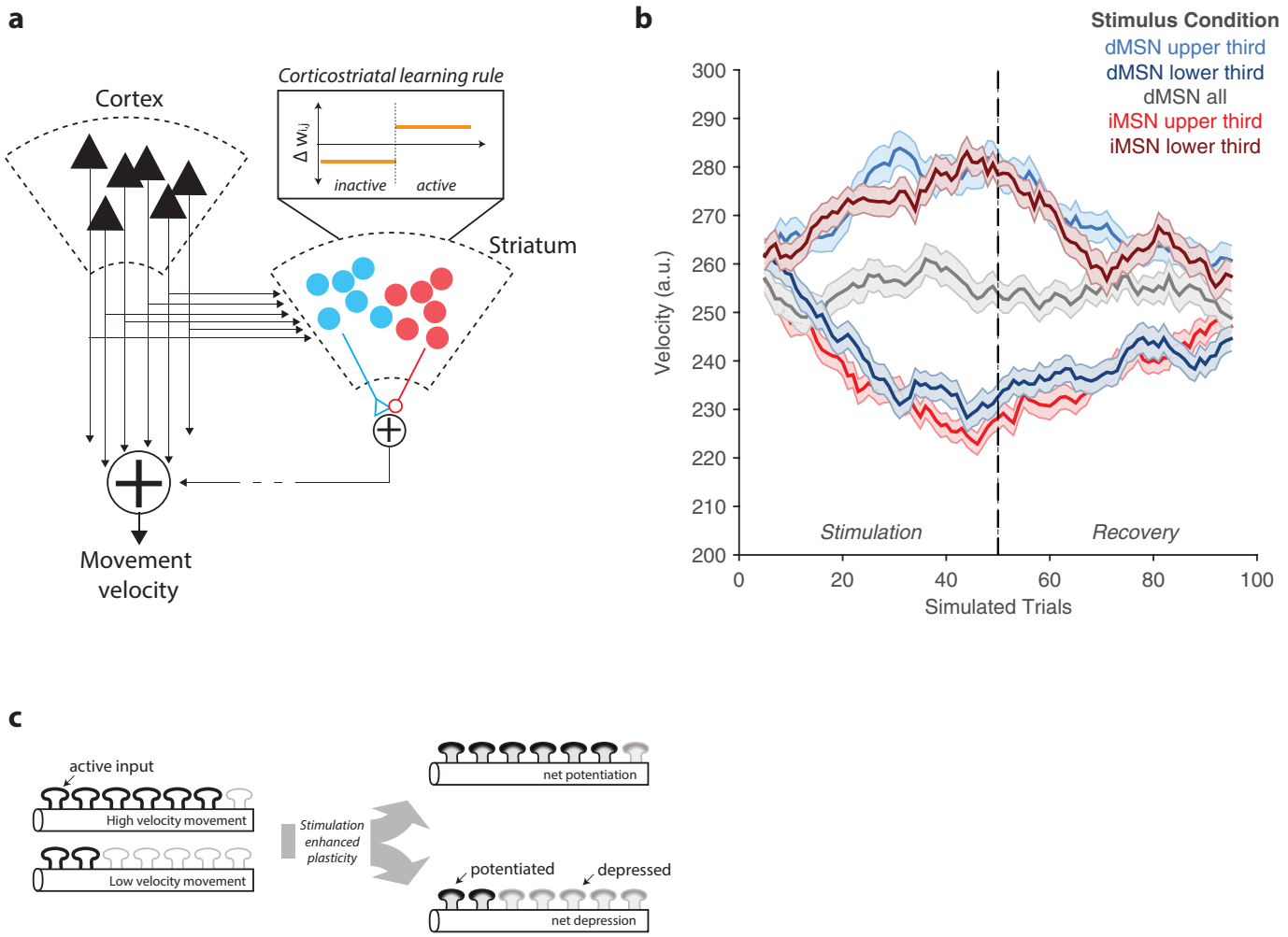
Extended Data Figure 4 | Variance of movement velocity does not change throughout a session. Each session was *z*-scored and the standard deviation for each movement number is plotted for dMSN (cyan) and iMSN (red) stimulation.

© 2016 Macmillan Publishers Limited. All rights reserved



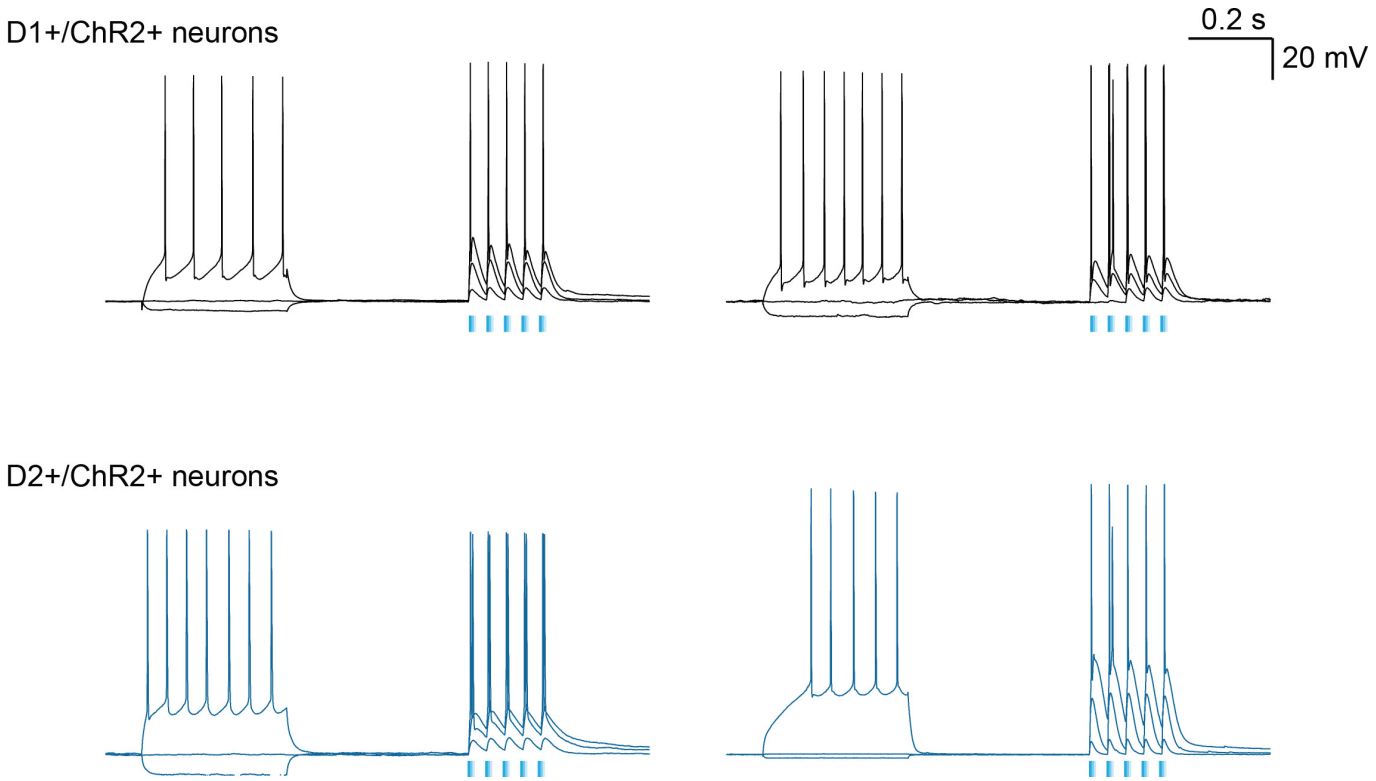
Extended Data Figure 5 | Non-selective stimulation does not affect motor control or initiation. **a**, **b**, 'All-stim' (**a**; 15 stimulation and 17 sham sessions from 4 dMSN mice; 17 stimulation and 20 sham sessions from 4 iMSN mice) and 'random stim' (**b**; 11 stimulation and 15 sham dMSN sessions from 3 dMSN mice, 8 stimulation and 12 sham sessions from 3 iMSN mice) summary data showing (top) the mean velocity as a function of movement number for dMSN (cyan) and iMSN (red) stimulation sessions and (bottom) histograms of the inter-move interval (IMI; left) and lick rate (LR; right). Shaded area indicates standard error of the mean. We found no differences between sham (black lines) and stimulation (coloured lines) sessions for either dMSN stimulation (cyan) or iMSN stimulation (red). **c**, Plot of average trajectory position aligned to stimulation onset for random dMSN (top) and iMSN stimulation (middle) for stimulation (coloured) and sham (black) sessions (in sham sessions, timing was randomly chosen, but no stimulation was given). Stimulation did not systematically induce forelimb movement. For reference, the bottom trace displays closed-loop dMSNs aligned to stimulation onset. Grey field represents the 450 ms stimulation period.

© 2016 Macmillan Publishers Limited. All rights reserved



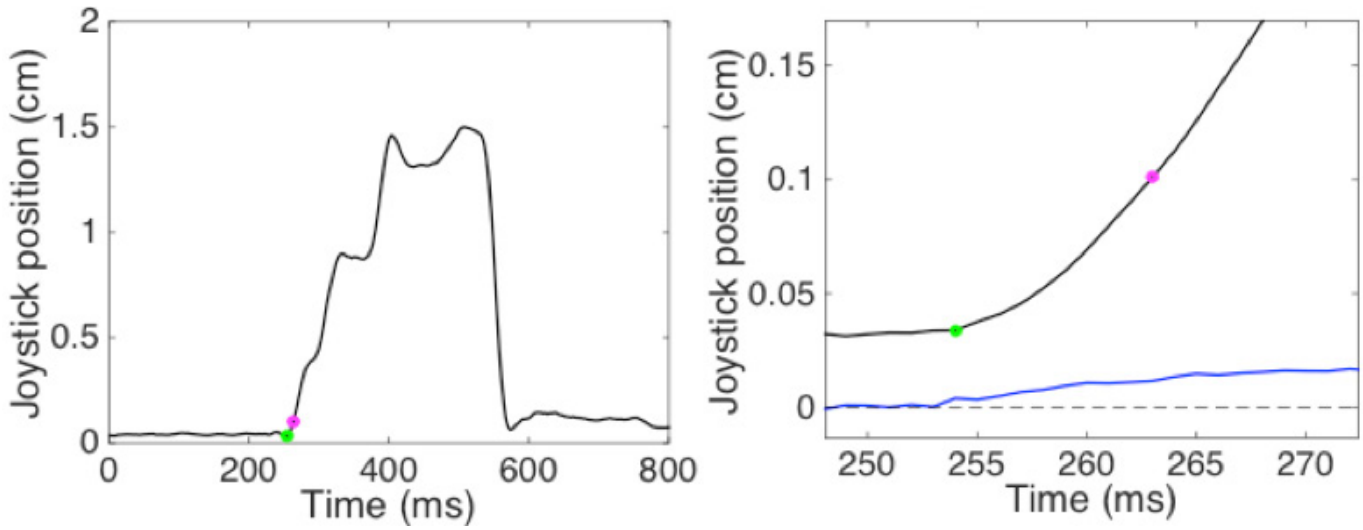
Extended Data Figure 6 | A corticostriatal circuit model that implements the MeSH learning rule. a, In the left panel, we present a schematic of the corticostriatal pathway consistent with known anatomical data[1]. Descending cortical outputs, largely from layer V of the neocortex, project subcortically and intracortically elaborating axon collaterals onto direct (blue) and indirect (red) MSNs in the dorsal striatum. By typical convention we assume that dMSNs have a net positive effect on behaviour (increase in velocity in this case) and iMSNs have a net inhibitory effect (decrease in velocity). These pathways are combined at the basal ganglia output nucleus (substantia nigra pars reticulata; not shown) and then combined with cortical drive to produce the net movement velocity. The model assumes that both dMSNs and iMSNs are positively correlated with cortical activity and with movement velocity. We assume a monotonic relationship between cortical activity and movement velocity[20]. The model is initialized at a presumptive steady-state in which weights between cortical inputs and dMSN and iMSN units are noisy, but distributed around 0.5 and bounded [0,1]. Most simulations were performed with 100 cortical units and 250 dMSNs and iMSNs each. Under all conditions weights are subject to updating according to a balanced plasticity rule (inset) in which inactive units are subject to depression and active units are subject to potentiation. All synapses drift back towards a mean of 0.5 thereby implementing a homeostatic set point to the weight distribution. Finally, photostimulation is assumed to enhance (90% increase) the magnitude of both depression and potentiation on stimulated trials in the stimulated population. Random sets of cortical inputs are assumed to be active on any given reach and are drawn from a gamma (shape parameters: 8, 63) distribution that gives a distribution of velocities similar to that observed experimentally. Further details of the model are provided in the Methods. **b**, Example simulations of 100 trials (first 50 receive stimulation according to conditions described in legend followed by 50 unstimulated recovery trials). Curves reflect averages and standard errors of 100 repetitions of the simulated condition. Other conventions as in main figures. **c**, Schematic of dendrite of MSN containing synapses active during limb movements. Synaptic plasticity enhanced by stimulation (inset in panel **a**) produces a net bias in synaptic weights when delivered in closed-loop. This bias can become uniform by permuting the active synapses on each simulated trial.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 7 | D1$^+$ and D2$^+$ MSNs can follow repetitive stimulus trains of ~20 Hz photostimulation *in vitro*. Example MSNs recorded *in vitro* in brain slices containing the DMS. Upper row are two example D1$^+$ dMSNs recorded from DMS of a Drd1a-Cre::Ai32 mouse. Lower row are two example D2$^+$ dMSN recorded from DMS of a Drd2a-Cre::Ai32 mouse. 3 example traces shown from each. Spiking is evoked by increasing current injection (traces selected for approximately similar evoked spiking rate) and ~500 ms later by a train of 5 pulses of blue light of increasing duty cycle. All cells recorded were able to follow rapid phasic stimulation and action potentials were reliably evoked on every stimulus of these 20 Hz trains (approximately similar to the stimulus trains used in stimulation experiments described in the text). Examples were selected from seven neurons from three D1$^+$ animals, and four neurons from three D2$^+$ animals that were recorded for this particular stimulation design.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 8 | Characterization of movement onset and reach initiation threshold crossing time. A sample reach from a sham block of 50 is shown, with eccentricity in black and velocity in blue (in right panel). The reach start is identified with the greed dot, the threshold crossing, 9 ms later when the reach eccentricity surpassed 0.1 cm, is identified with the magenta dot. The beginning of the reach (green) was assessed offline for each reach and was determined to be the first time point, sampled at 1 kHz, constituting the increasing velocity associated with that reach.

© 2016 Macmillan Publishers Limited. All rights reserved

Extended Data Table 1 | Lick and move rates from each animal for sham upper third stimulation sessions

dMSN lick frequency (Hz):	Sham (sem)		Stim (sem)
Mouse 1	7.2 (0.3)		7.4 (0.3)
Mouse 2	7.0 (0.1)		6.9 (0.1)
Mouse 3	7.2 (0.1)		7.2 (0.2)
Mouse 4	7.3 (0.1)		7.2 (0.1)
iMSN lick frequency (Hz):			
Mouse 5	7.1 (0.6)		7.4 (0.3)
Mouse 6	6.9 (0.9)		6.6 (0.3)
Mouse 7	6.8 (0.1)		7.1 (0.2)
Mouse 8	6.7 (0.1)		6.8 (0.1)
dMSN inter-move interval (seconds):			
Mouse 1	5.9 (0.6)		5.6 (0.5)
Mouse 2	6.2 (0.3)		5.9 (0.3)
Mouse 3	6.1 (0.4)		6.8 (0.2)
Mouse 4	5.7 (0.3)		6.7 (0.4)
iMSN inter-move interval (seconds):			
Mouse 5	7.6 (0.4)		7.3 (0.4)
Mouse 6	5.9 (0.3)		6.2 (0.3)
Mouse 7	6.2 (0.4)		6.8 (1.1)
Mouse 8	6.0 (0.3)		6.3 (0.4)

Increases in the mean for lick frequency indicate a faster rate of licking (more motor output), while increases in the mean for inter-move interval indicate a slower rate of licking (less motor output). Lick rate changes per animal were not significant ($P > 0.4$ for all but one iMSN animal, which exhibited a paradoxically faster lick rate, $P = 0.18$). Inter-move interval changes were also not significant within each animal ($P > 0.3$ except one dMSN animal which paradoxically showed an increased inter-move interval, $P = 0.15$). Mouse number is only meant as an index and does not reflect timing or order of experiments, most of which were done in the same period of time.

© 2016 Macmillan Publishers Limited. All rights reserved

LETTER

doi:10.1038/nature17972

FGF8 and SHH substitute for anterior–posterior tissue interactions to induce limb regeneration

Eugeniu Nacu[1,2], Elena Gromberg[1,2], Catarina R. Oliveira[1,3], David Drechsel[2] & Elly M. Tanaka[1,2]

In salamanders, grafting of a left limb blastema onto a right limb stump yields regeneration of three limbs, the normal limb and two 'supernumerary' limbs[1–4]. This experiment and other research have shown that the juxtaposition of anterior and posterior limb tissue plus innervation are necessary and sufficient to induce complete limb regeneration in salamanders[5–10]. However, the cellular and molecular basis of the requirement for anterior–posterior tissue interactions were unknown. Here we have clarified the molecular basis of the requirement for both anterior and posterior tissue during limb regeneration and supernumerary limb formation in axolotls (*Ambystoma mexicanum*). We show that the two tissues provide complementary cross-inductive signals that are required for limb outgrowth. A blastema composed solely of anterior tissue normally regresses rather than forming a limb, but activation of hedgehog (HH) signalling was sufficient to drive regeneration of an anterior blastema to completion owing to its ability to maintain fibroblast growth factor (FGF) expression, the key signalling activity responsible for blastema outgrowth. In blastemas composed solely of posterior tissue, HH signalling was not sufficient to drive regeneration; however, ectopic expression of FGF8 together with endogenous HH signalling was sufficient. In axolotls, FGF8 is expressed only in the anterior mesenchyme and maintenance of its expression depends on sonic hedgehog (SHH) signalling from posterior tissue. Together, our findings identify key anteriorly and posteriorly localized signals that promote limb regeneration and show that these single factors are sufficient to drive non-regenerating blastemas to complete regeneration with full elaboration of skeletal elements.

Limb regeneration in salamanders, including axolotls, proceeds in four steps: injury, blastema formation, sustained blastema outgrowth, and differentiation. Previous studies have shown that blastema formation is dependent on the presence of nerves at the injury site, and that sustained outgrowth requires the presence of both anterior and posterior limb cells[5–10]. Remarkably, despite the large cadre of molecular information that has accumulated on limb patterning in the past decades, the cellular and molecular logic underlying the requirement for anterior and posterior cell juxtaposition in maintaining limb regeneration is unknown. Two models relying on different cellular principles have been proposed to explain this requirement. In the polar coordinate model, a graded spectrum of positional values around the limb circumference manifests in the form of cell surface properties. Disruption of this gradation as a result of transplantation or amputation would be recognized by cells at the border, stimulating them to proliferate and generate daughter cells that intercalate the missing circumferential identities and thereby resulting in distal outgrowth of the limb tissue[11]. By contrast, the boundary model subdivides the limb circumference into four domains: anterior–dorsal, anterior–ventral, posterior–dorsal, and posterior–ventral, and proposes that the intersection between the domains is required to induce the posterior domains to express a secreted morphogen that is required for growth and patterning of the limb[12]. Thus, complementing, cross-inductive interactions between domains rather than local detection of differences form the conceptual basis of the boundary model.

Anterior-only blastemas (ABs) composed solely of anterior tissue can be created by deviating nerve endings to an anterior-facing lateral wound in the upper limb[8,9]. A mid-bud blastema forms at the lateral site but eventually regresses, unless a patch of posterior limb skin (including dermal mesenchyme) is grafted to the anterior site; the graft allows the blastema to grow and form a fully patterned limb, called an accessory limb. We first searched for posterior-associated molecules that could stimulate complete limb regeneration from ABs and found that stimulation of HH signalling via the global application of smoothened agonist (SAG) was sufficient to prevent regression of an AB and to elicit regeneration of a limb with clear skeletal patterning along the proximo–distal axis and the full complement of constituent limb cell types (Fig. 1a, Extended Data Figs 1 and 2a, b, d and Supplementary Table 1). The accessory limbs displayed varying digit numbers that depended on SAG concentration (Extended Data Fig. 2a). We corroborated these findings by overexpressing human SHH in ABs via baculovirus-mediated delivery; this also led to patterned limb outgrowth ($n = 8$ of 32; Fig. 1b and Supplementary Table 1). These results show that activation of the HH pathway is sufficient to induce regeneration from an anterior-only blastema. Conversely, inhibition of HH signalling by global application of cyclopamine blocked accessory limb outgrowth from ABs complemented with a graft of posterior skin (Extended Data Fig. 2c), indicating that HH signalling is required for ectopic blastema outgrowth when a full complement of positional identities is present.

To gain an insight into the mechanism of SAG-induced outgrowth of ABs, we investigated whether HH signalling induced a stable change from an anterior to a posterior determination state in parts of the ectopic limb tissue, or whether SAG treatment stimulated anteriorly determined cells to grow out a limb. Reamputation of SAG-induced accessory limbs (SAG-ALs) yielded either no outgrowth ($n = 2$ of 5) or outgrowth of a spike ($n = 3$ of 5), suggesting that a stable anterior–posterior discontinuity that could induce a secondary limb had not been established (Extended Data Fig. 3a). Transplantation is the cleanest way to test determination state. Therefore we transplanted skin from SAG-ALs to innervated anterior or posterior wounds and looked for secondary accessory limb formation (Extended Data Fig. 3b). Limb outgrowth occurred when the SAG-AL skin was transplanted to posterior ($n = 15$ of 21) but not anterior wounds ($n = 0$ of 22), indicating that the SAG-AL limb tissue retains an anterior rather than posterior determination state (Extended Data Fig. 3c, d and Supplementary Table 2).

These results suggested that SAG may promote accessory limb outgrowth by upregulating a downstream growth-stimulating signalling loop that is involved in limb development; namely, posteriorly secreted SHH that stimulates upregulation of *gremlin1* (*grem1*) and consequently *fgf* genes, whose protein products in turn maintain *shh*

[1]DFG Research Center for Regenerative Therapies, Technische Universität Dresden, 01307 Dresden, Germany. [2]Max Planck Institute of Molecular Cell Biology and Genetics, 01307 Dresden, Germany. [3]Graduate Program in Areas of Basic and Applied Biology, Abel Salazar Biomedical Sciences Institute, University of Porto, 4099-003 Porto, Portugal.

© 2016 Macmillan Publishers Limited. All rights reserved

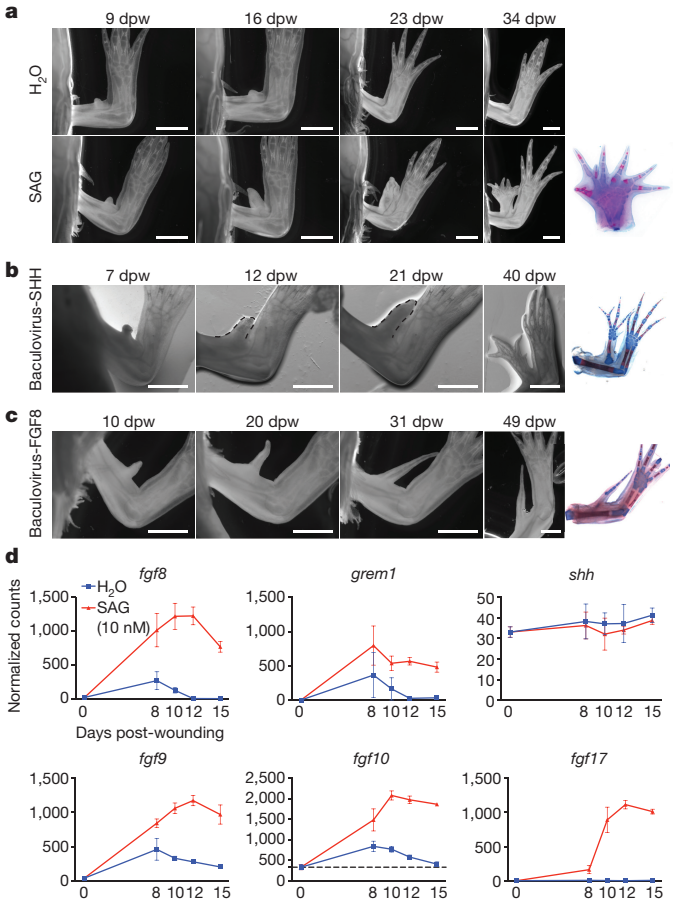


Figure 1 | Activation of HH signalling or expression of FGF8 in anterior blastemas (ABs) is sufficient to drive outgrowth of accessory limbs. **a**, Accessory limbs develop from ABs treated with 10 nM SAG ($n = 43$ of 66), but not from ABs treated with water ($n = 0$ of 22). Alcian blue/alizarin red staining reveals the skeletal pattern with supernumerary digits and carpal elements in the accessory limb. Data from three experiments. See Supplementary Table 1 for details. **b**, Ectopic expression of SHH in ABs by baculovirus transduction is sufficient to drive accessory limb outgrowth ($n = 8$ of 32, data from two experiments). Black dashed lines demarcate the outgrowing accessory limb. See Supplementary Table 1 for details. **c**, Ectopic expression of FGF8 in ABs by baculovirus transduction is sufficient to drive accessory limb outgrowth ($n = 14$ of 57, data from four experiments). Alcian blue/alizarin red staining reveals the skeletal pattern with a single digit in the accessory limb. See Supplementary Table 4 for details. **d**, Expression of FGF–SHH loop components in ABs untreated or treated with 10 nM SAG and analysed with Nanostring nCounter technology. Each data point is the mean normalized count number from biological replicates ($n = 3$ or 4; for individual n values and complete data set, see Supplementary Table 7a; data from one experiment). Error bars represent s.d. For data points where the s.d. was low, no error bars are shown. The dashed line is for visual aid and indicates the level of mature (day 0) expression. For *in situ* hybridization of selected genes, see Extended Data Fig. 4. Scale bars, 2 mm. dpw, days post-wounding. For alcian blue/alizarin red staining, cartilage is blue and ossifications are red.

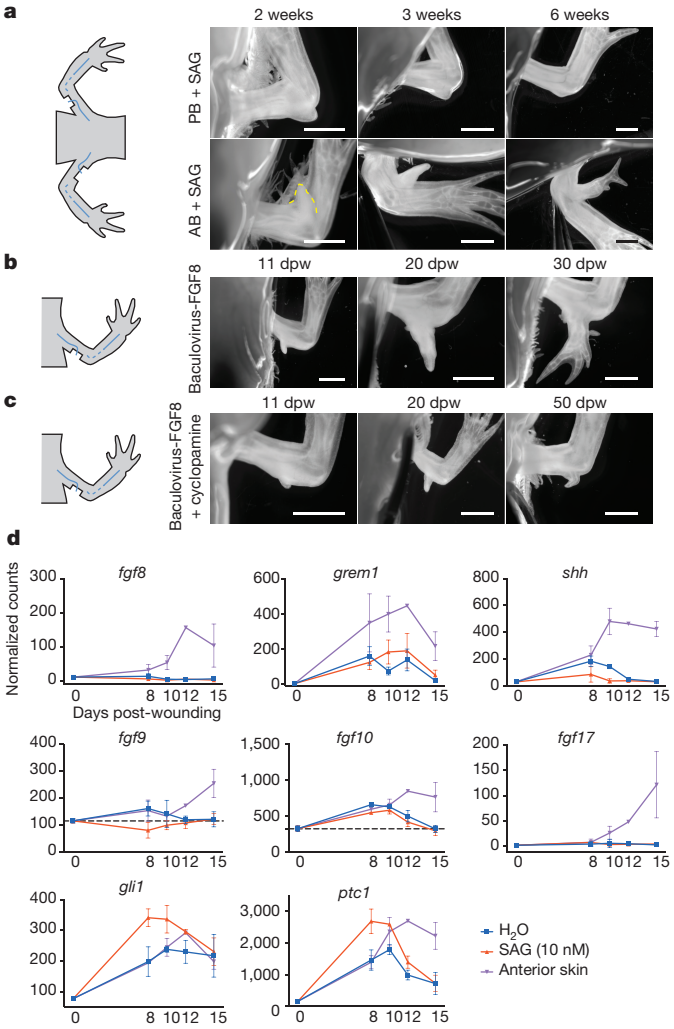


Figure 2 | FGF8 is not detected in posterior blastemas (PBs) and its ectopic expression is sufficient to drive accessory limb outgrowth in the presence of intact endogenous HH signalling. **a**, Accessory limbs did not grow from PBs (upper row) of animals treated with 10 nM SAG ($n = 1$ of 32), while ABs on the contralateral forelimbs of the same animals (lower row) developed into accessory limbs ($n = 23$ of 32), indicating efficient deactivation of HH signalling. Data from one experiment. Yellow dashed line demarcates the AB. See Extended Data Fig. 6a–c, f for controls and Supplementary Table 1 for details. **b**, PBs expressing FGF8 after baculovirus transduction grow into accessory limbs ($n = 27$ of 50; data from three experiments). See Extended Data Fig. 6d–e, g for controls and Supplementary Table 5 for details. **c**, PBs expressing ectopic FGF8 and treated with 8 μM cyclopamine, an inhibitor of HH signalling, do not develop into accessory limbs ($n = 3$ of 48; data from four experiments). See Extended Data Fig. 10 for controls and Supplementary Table 6 for details. **d**, Expression of FGF–SHH loop components in PBs treated with 10 nM SAG, untreated or complemented with anterior skin, analysed with Nanostring nCounter technology. Note the lack of *fgf8* induction by SAG despite the induction of the HH signalling targets *gli1* and *ptc1*, which indicates functional HH signalling. Each data point is the mean normalized count from biological replicates ($n = 2$ or 3; for individual n values and complete data set, see Supplementary Table 7b; data from one experiment). Error bars represent s.d. For data points where the s.d. was low or where $n = 2$, no error bars are shown. The dashed lines are for visual aid and indicate the level of mature (day 0) expression. Extended Data Fig. 7 shows *in situ* hybridization of selected genes. Scale bars, 2 mm. dpw, days post-wounding.

expression in posterior cells[13–25]. We therefore quantified the expression of key limb development factors and limb blastema genes in SAG-treated versus untreated ABs by Nanostring nCounter analysis (a method for counting RNA transcripts) and by RNA *in situ* hybridization. As expected, we observed upregulation of the blastema-associated transcription factors *prrx1*, *msx1*, *msx2* and *twist1* and of the anterior marker *gli3* in ABs treated with SAG or water (Extended Data Fig. 4a, b and Supplementary Table 7a). Interestingly, a set of extracellular signalling factors that included *fgf8*, *fgf9*, *fgf10* and *grem1* were sustainably and highly upregulated in SAG-treated ABs but showed only an initial, unsustained upregulation in untreated ABs. *Fgf17* was expressed only in ABs treated with SAG, and *shh* was not expressed in ABs at all (Fig. 1d, Extended Data Figs 4 and 8 and Supplementary Tables 3, 7a and 8).

© 2016 Macmillan Publishers Limited. All rights reserved

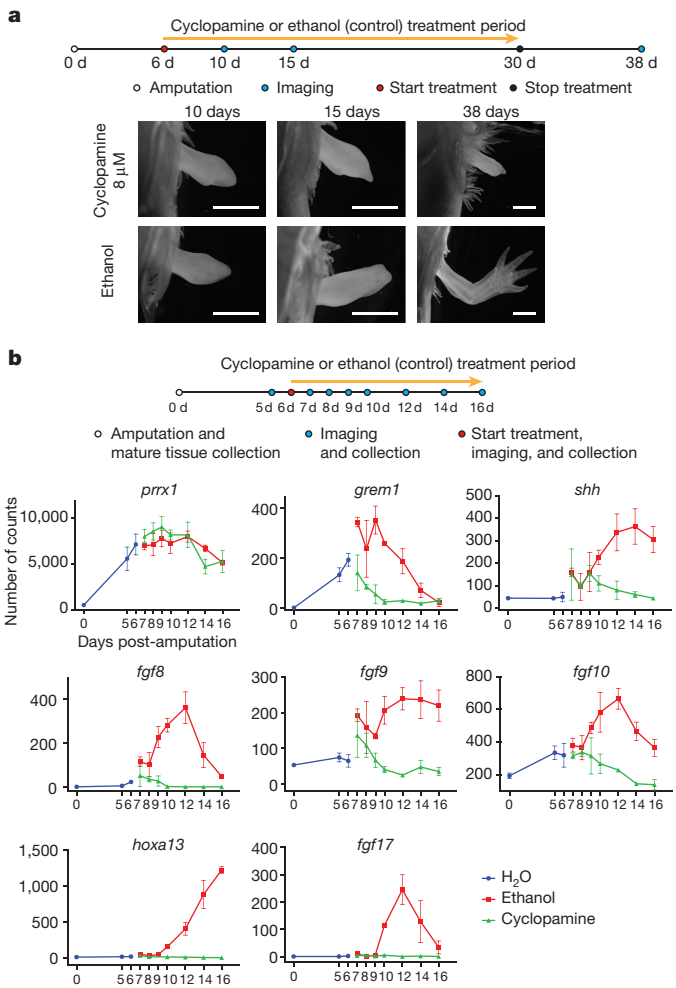


Figure 3 | HH signalling is required for maintaining blastema outgrowth and regeneration. a, Inhibition of HH signalling with cyclopamine blocks progression of limb regeneration. Limbs treated with cyclopamine (8 μM) starting from day 6 after amputation do not regenerate ($n = 0$ of 14), whereas ethanol-treated controls do ($n = 6$ of 6). Data from three experiments. Scale bars, 2 mm. **b**, Nanostring nCounter gene expression analysis of limbs treated with 8 μM cyclopamine or ethanol. The schema illustrates the experimental setup. The blastema gene *prrx1* shows a similar profile in control and cyclopamine-treated samples, whereas genes involved in the SHH–FGF loop and the late genes *fgf17* and *hox13* diverge under the two conditions. Each data point is the mean normalized count of data from biological replicates ($n = 3$ or 4; for complete data set and individual n values, see Supplementary Table 10; data from one experiment). Error bars represent s.d. For data points where the s.d. was low, no error bars are shown.

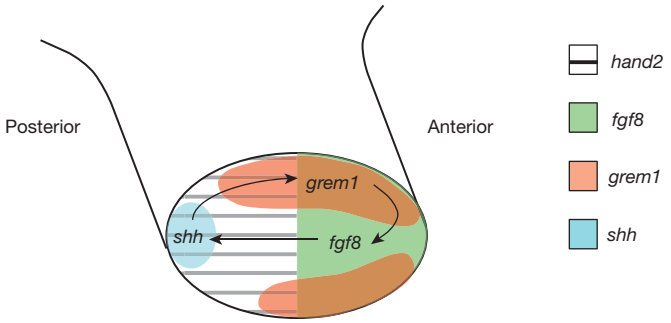


Figure 4 | Molecular circuitry of signals involved in anterior and posterior tissue requirement of regeneration. Anterior cells of the blastema express *fgf8*, whereas *shh* is expressed in the most posterior part. The initial upregulation of *fgf8* in anterior cells is independent of interaction with posterior cells and the initial upregulation of *shh* in posterior cells is independent of interaction with anterior cells. Later, the expression of *fgf8* and *shh* becomes interdependent in a positive feedback loop, probably involving GREM1, that is expressed primarily in the anterior half of the blastema but also in some posterior *hand2*-expressing cells. FGFs are the factors that can drive blastema outgrowth and limb regeneration.

These data pointed to FGF signalling as a potentially critical effector of SAG in ABs. To test this theory, SAG-treated ABs were exposed to the FGFR1/FGFR3 signalling inhibitor PD173074 starting at day 10 after wounding. Treatment of ABs with SAG alone induced accessory limb formation ($n = 14$ of 14), whereas addition of PD173074 blocked SAG-induced accessory limb outgrowth ($n = 0$ of 16) (Extended Data Fig. 5a). To test whether FGF signalling can drive accessory limb outgrowth, FGF8 was ectopically expressed in ABs by baculovirus transduction; this yielded single digit outgrowths (Fig. 1c, Extended Data Fig. 5b and Supplementary Table 4) without radius or ulna formation or anterior–posterior digit elaboration. These data show that FGF signalling is both necessary and sufficient to drive accessory limb outgrowth from ABs.

We next tested whether SAG treatment was sufficient to induce outgrowth of posterior-only blastemas (PBs). PBs also regress unless complemented with an anterior skin graft. Interestingly, we found that SAG did not induce PBs to form accessory limbs ($n = 1$ of 48) (Fig. 2a, Extended Data Fig. 6a–c, f and Supplementary Table 1). Nanostring analysis (Fig. 2d and Supplementary Tables 7b, 9) and RNA *in situ* hybridization (Extended Data Fig. 7 and Supplementary Table 3) for *gli1*, *ptc1* and *shh* revealed that these factors were expressed in PBs with or without SAG, indicating that the HH signalling system is intact in PBs but is not sufficient to induce regeneration in this context. In addition, *grem1* and *fgf10* were also initially upregulated in PBs with or without SAG, and *fgf9* was initially upregulated in untreated PBs. On the other hand, *fgf8* and *fgf17* were not upregulated in PBs with or without SAG (Fig. 2d, Extended Data Figs 7 and 8 and Supplementary Tables 3, 7b, 9). RNA *in situ* hybridization on normal blastemas revealed that the *fgf8* signal localized to the anterior half of the blastema, being excluded from the posterior *hand2*-expressing zone. Interestingly *grem1, fgf9, fgf10* and *fgf17* were not excluded from the posterior *hand2*-expressing domain (Extended Data Fig. 9a, b). These results suggested that FGF8 is the anteriorly expressed FGF that is limiting in PB outgrowth. To test whether FGF8 expression could substitute for anterior cells in PBs to drive accessory limb outgrowth, we transduced the posterior regions of axolotl limbs with baculovirus expressing *fgf8* and created PBs several days later. FGF8 expression alone was sufficient to induce accessory limb outgrowth in 54% of cases ($n = 27$ of 50), whereas no significant outgrowth was observed after control mCherry expression (Fig. 2b, Extended Data Fig. 6d, e, g and Supplementary Table 5). Interestingly, FGF8-mediated outgrowth from PBs required endogenous HH signalling, as co-treatment with cyclopamine blocked accessory limb outgrowth (Fig. 2c, Extended Data Fig. 10 and Supplementary Table 6). Our results show that FGF8 expression together with endogenous HH signalling is sufficient to drive accessory limb outgrowth from PBs.

We next investigated the relevance of our findings to normal regeneration. Treatment of upper arm blastemas with cyclopamine starting at day 6 after amputation blocked regeneration (Fig. 3a) and yielded changes in the expression of extracellular signalling factors (Fig. 3b and Supplementary Table 10). The limb blastema marker *prrx1* showed similar profiles in blastemas treated with cyclopamine or ethanol (control) up to day 16. In contrast, *fgf8, fgf9,* and *fgf10* showed sustained expression under control conditions, whereas in cyclopamine-treated blastemas, the initial induction of *fgf8, fgf9,* and *fgf10* up to day 8 was followed by downregulation. Interestingly, *grem1* showed a faster kinetic, peaking after 6 days before being downregulated in cyclopamine-treated blastemas, and *shh* expression started to be downregulated at day 10 in cyclopamine-treated blastemas. These

© 2016 Macmillan Publishers Limited. All rights reserved

profiles are consistent with GREM1 being upstream of *fgf* gene expression, and with FGFs being required for *shh* maintenance[14,19,22–24]. Two late-stage genes associated with distal limb bud development, *fgf17* and *hoxa13*, were strongly expressed only in ethanol-treated blastemas, suggesting that FGF–SHH signalling is required for substantial distalization during limb regeneration. These results show that the FGF–SHH loop is active during regeneration and is required for blastema outgrowth. Additionally, the anterior localization of *fgf8*, but not of *fgf9*, *fgf10* or *fgf17*, in the limb blastema, and the lack of expression of *fgf8* in PB cells, suggest that the cross-induction between anteriorly restricted *fgf8* and posteriorly restricted *shh* (Extended Data Fig. 9a, c) is an essential part of normal limb regeneration (Fig. 4).

In summary, we have delineated the mechanism underlying the anterior–posterior juxtaposition that has long been known to be required for limb regeneration. We have identified single factors that can substitute for anterior and posterior tissue functions: SHH for posterior and FGF8 for anterior tissue. This work shows that, in axolotls, the anterior and posterior tissue provide cells that are restricted in their competence to express complementary soluble signalling factors required for the growth and patterning of a mid-bud blastema, a result consistent with the boundary model of limb regeneration[12]. Posteriorly localized SHH is required for the maintenance of anteriorly restricted FGF8 expression and of non-restricted expression of FGF9, FGF10 and FGF17. Sustained FGF signalling is the key limiting factor for extended proliferation of anterior and posterior cells and subsequent morphogenesis (Fig. 4).

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 15 June 2014; accepted 5 April 2016.
Published online 27 April 2016.

1. Iten, L. E. & Bryant, S. V. The interaction between the blastema and stump in the establishment of the anterior–posterior and proximal–distal organization of the limb regenerate. *Dev. Biol.* **44,** 119–147 (1975).
2. Maden, M. Structure of supernumerary limbs. *Nature* **286,** 803–805 (1980).
3. Maden, M. Supernumerary limbs in amphibians. *Integr. Comp. Biol.* **22,** 131–142 (1982).
4. Stocum, D. L. Determination of axial polarity in the urodele limb regeneration blastema. *J. Embryol. Exp. Morphol.* **71,** 193–214 (1982).
5. Bryant, S. V. Regenerative failure of double half limbs in *Notophthalmus viridescens*. *Nature* **263,** 676–679 (1976).
6. Bryant, S. V. & Baca, B. A. Regenerative ability of double-half and half upper arms in the newt, *Notophthalmus viridescens. J. Exp. Zool.* **204,** 307–323 (1978).
7. Stocum, D. L. Regeneration of symmetrical hindlimbs in larval salamanders. *Science* **200,** 790–793 (1978).
8. Endo, T., Bryant, S. V. & Gardiner, D. M. A stepwise model system for limb regeneration. *Dev. Biol.* **270,** 135–145 (2004).
9. Satoh, A., Gardiner, D. M., Bryant, S. V. & Endo, T. Nerve-induced ectopic limb blastemas in the axolotl are equivalent to amputation-induced blastemas. *Dev. Biol.* **312,** 231–244 (2007).
10. Nacu, E. & Tanaka, E. M. Limb regeneration: a new development? *Annu. Rev. Cell Dev. Biol.* **27,** 409–440 (2011).
11. Bryant, S. V., French, V. & Bryant, P. J. Distal regeneration and symmetry. *Science* **212,** 993–1002 (1981).
12. Meinhardt, H. A boundary model for pattern formation in vertebrate limbs. *J. Embryol. Exp. Morphol.* **76,** 115–137 (1983).
13. Laufer, E., Nelson, C. E., Johnson, R. L., Morgan, B. A. & Tabin, C. Sonic hedgehog and Fgf-4 act through a signaling cascade and feedback loop to integrate growth and patterning of the developing limb bud. *Cell* **79,** 993–1003 (1994).
14. Chiang, C. *et al.* Manifestation of the limb prepattern: limb development in the absence of sonic hedgehog function. *Dev. Biol.* **236,** 421–435 (2001).
15. Harfe, B. D. *et al.* Evidence for an expansion-based temporal Shh gradient in specifying vertebrate digit identities. *Cell* **118,** 517–528 (2004).
16. Towers, M., Mahood, R., Yin, Y. & Tickle, C. Integration of growth and specification in chick wing digit-patterning. *Nature* **452,** 882–886 (2008).
17. Zhu, J. *et al.* Uncoupling Sonic hedgehog control of pattern and expansion of the developing limb bud. *Dev. Cell* **14,** 624–632 (2008).
18. Niswander, L., Jeffrey, S., Martin, G. R. & Tickle, C. A positive feedback loop coordinates growth and patterning in the vertebrate limb. *Nature* **371,** 609–612 (1994).
19. Zúñiga, A., Haramis, A. P., McMahon, A. P. & Zeller, R. Signal relay by BMP antagonism controls the SHH/FGF4 feedback loop in vertebrate limb buds. *Nature* **401,** 598–602 (1999).
20. Lewandoski, M., Sun, X. & Martin, G. R. Fgf8 signalling from the AER is essential for normal limb development. *Nature Genet.* **26,** 460–463 (2000).
21. Moon, A. M. & Capecchi, M. R. Fgf8 is required for outgrowth and patterning of the limbs. *Nature Genet.* **26,** 455–459 (2000).
22. Khokha, M. K., Hsu, D., Brunet, L. J., Dionne, M. S. & Harland, R. M. Gremlin is the BMP antagonist required for maintenance of Shh and Fgf signals during limb patterning. *Nature Genet.* **34,** 303–307 (2003).
23. Panman, L. *et al.* Differential regulation of gene expression in the digit forming area of the mouse limb bud by SHH and gremlin 1/FGF-mediated epithelial-mesenchymal signalling. *Development* **133,** 3419–3428 (2006).
24. Nissim, S., Hasso, S. M., Fallon, J. F. & Tabin, C. J. Regulation of Gremlin expression in the posterior limb bud. *Dev. Biol.* **299,** 12–21 (2006).
25. Mariani, F. V., Ahn, C. P. & Martin, G. R. Genetic evidence that FGFs have an instructive role in limb proximal-distal patterning. *Nature* **453,** 401–405 (2008).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank B. Gruhl, S. Moegel, M. Armstead, H. Goers, A. Wagner and S. Kaudel for animal care, H. Q. Le and K. Goehler for assistance, C. Cannistraci for advice on statistical analysis, C. Antos, D. Knapp, W. Masselink, T. Sugiura and Y. Taniguchi for advice on the manuscript, K. Crawford for discussion, S. Eaton for providing us with the plasmid with human *SHH*, K. Airenne, S. Ylä-Herttuala, and M. Kaikonnen for the plasmid with VSV-GED, members of the MPI-CBG protein facility for assistance in baculovirus preparation, and members of the MPI-CBG antibody facility for the anti-MHC antibody. This work was supported by central funds from the CRTD and MPI-CBG and an ERC Advanced Investigator grant to E.M.T. C.R.O. was supported by the Portuguese Foundation for Science and Technology (FCT). Schematic illustrations of axolotl limbs were redrawn based on illustrations from ref. 10 under copyright transfer agreement to E.N. and E.M.T. as authors.

Author Contributions E.N. and E.M.T. conceived the project, conceptualized the experiments and wrote the manuscript. E.N. designed and managed the experiments. E.N. and E.G. performed most of the limb experiments and *in situ* hybridizations. C.R.O. and D.D. established the baculovirus transduction system and C.R.O. performed the human *SHH* transductions. E.N., E.G., C.R.O. and E.M.T. analysed data. Figures and tables were prepared by E.N. and E.G.

Author Information The following sequences have been deposited in the NCBI GenBank database with the corresponding accession numbers: *fgf10*, KU882013; *fgf17*, KU882014; *grem1*, KU882015; *hand2*, KU882016; *gli3*, KU882017; *hoxd13*, KU882018; *fgf8*, KU882019; *fgf9*, KU882020. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.N. (eugeniu_nacu@harvard.edu) or E.M.T. (elly.tanaka@crt-dresden.de).

© 2016 Macmillan Publishers Limited. All rights reserved

METHODS

Data reporting. Animal distribution to experimental groups was randomized within the selection criteria outlined below for different experiments. The investigators were not blinded to allocation during experiments and outcome assessment.

Animal procedures. Axolotls (*Ambystoma mexicanum*) used in all experiments were bred in captivity in our facility, maintained at 17–24 °C in tap water and fed with artemia. Axolotls of the leucistic (d/d) strain were used in all experiments. Non-transgenic animals were used in all experiments unless indicated otherwise. When necessary, transgenic animals of the *caggs::egfp* genotype were used[26].

Animals of length between 2 and 4 cm from snout to cloaca (that is, between 4 and 8 cm from snout to tail tip) were used in all experiments, except for the SAG treatment experiment in which larger 13–14-cm (snout to tail tip) animals were used (Extended Data Fig. 2b). In any given experiment, the range in the length (snout to cloaca) of animals used was no more than 0.4 cm, to ensure comparable dynamics of regeneration. Among different replicate experiments, animals of similar size were used. For any given experiment, sibling animals were preferentially used. In cases when this was not possible, animals from different matings were evenly distributed among the different experimental subgroups. It was not possible to determine the sex of the animals at the stage at which the experiment was performed and therefore this was not taken into account.

All animal procedures were carried out in accordance with the laws and regulations of the State of Saxony, Germany.

Surgical procedures. For all experimental procedures, axolotls were anaesthetized in 0.03% benzocaine (Sigma) before surgery. All surgeries were performed using Olympus SZX10 microscopes.

To initiate normal limb regeneration, limbs were amputated either through the upper arm or lower arm.

To generate ABs and PBs, wounds were created by cutting out a rectangular piece of skin from the anterior or posterior side of the distal half of the upper arm. Two ventral nerves were dissected free, severed at the elbow level and deflected to the wound.

Skin transplantation to wounds was performed similarly. A slightly larger patch of skin compared to AB or PB production was removed from the host distal half of the upper arm and nerves were deflected to the wound. A square skin flap from the donor arm was placed on the distal half of the host wound, while leaving the proximal half uncovered with nerves sticking out. To allow healing of the transplant, animals were left for 15 min to 1 h in a Petri dish humidified with 0.03% benzocaine; afterwards animals were transferred into tap water.

Treatment of animals with chemical agonists or inhibitors. In all experiments that required animals to be treated with chemical compounds, the compounds were dissolved in tap water and animals were kept in that solution. Solutions were made freshly and exchanged daily. The chemical compounds used in the study were: smoothened agonist SAG (EMD Millipore), PD173074 (Tocris Bioscience), cyclopamine (LC labs), dimethyl sulfoxide DMSO (Sigma) and ethanol (99.9% VWR chemicals). Chemical solutions were prepared in glass bottles covered with aluminium foil that protected the solutions from light, and stirred for a minimum of 30 min with a magnetic stirrer. Solutions were distributed to the boxes in which animals were kept. The animals were then transferred into the boxes with solutions and kept in darkness. During treatment, animals were fed daily for 30 min to 1 h under lab light or preferentially in darkness. Feeding was started 30 min to 1 h before the chemical solution was exchanged.

SAG treatment of ABs and PBs. ABs and PBs were generated by surgical methods as described above. Treatment with SAG was performed as described above and started immediately after surgery or 3 days after surgery for some PB samples. The duration of treatment with SAG varied; the duration of treatment for each experiment is indicated in Supplementary Table 1.

Cyclopamine treatment of blastemas, ABs complemented with posterior skin, and PBs complemented with anterior skin. Limbs were amputated or injured as above. An 8 μM cyclopamine solution in tap water was prepared from a stock of 4 mM cyclopamine in ethanol as described. We found that the effect of cyclopamine was dependent not only on the final concentration but also on the volume of the solution in which animals were kept. In all experiments, the volume of solution per animal was 60 ml. A control solution of ethanol in tap water with similar dilution and volume was also prepared. Treatment with solutions was started 6 days after surgery. The duration of treatment is indicated in the experimental schemas associated with each experiment.

Exogenous expression of genes in ABs or PBs by baculovirus transduction. Baculovirus technology[27–29] was established in our lab for use as a transient expression system in the axolotl and will be described elsewhere. Baculovirus was pseudotyped with *vsv-ged* gene, which was inserted into the rescue vector under the baculovirus polyhedrin promoter. Genes of interest (human *SHH*, axolotl *fgf8*, and *mCherry*) were cloned into the baculovirus rescue vector under the *CMV* promoter using standard restriction enzyme methods.

The generation of pseudotyped baculoviruses carrying the gene of interest was carried out by co-transfection of the rescue vector together with replication-incompetent baculovirus DNA into a modified *Spodoptera frugiperda* cell line (expresSF+, ProteinSciences Corporation). Upon culture expansion, recombinant baculoviruses were collected, concentrated and purified. Baculovirus titre was assessed by end-point dilution assay in SF-9 Easy Titer cells, a modified *Spodoptera frugiperda* cell line expressing a reporter *egfp* gene under the control of the baculovirus polyhedrin promoter[30].

Expression of SHH in ABs by baculovirus transduction was achieved by injecting baculovirus–SHH into the anterior sides of mature upper arms, followed by surgery 5 days after injection. Titres of injected baculovirus suspension are shown in Supplementary Table 1.

Expression of mCherry or FGF8 was achieved by injecting the corresponding baculovirus into the posterior or anterior sides of mature upper arms. Titres of the viral suspensions injected were: mCherry, 2.7×10^{10} or 9.4×10^{10}; FGF8, 2.7×10^{10} or 8.5×10^{10}.

Injections were performed using a glass capillary needle pulled on a micropipette puller (Sutter Instrument Co. model P-97). Glass capillaries (1.2 mm OD, 0.9 mm ID borosilicate glass) were acquired from Harvard Apparatus. For injection into the anterior upper arm, either a tiny incision was made with surgical scissors (FineScienceTools No 15024-10) or the skin was punctured with a 30G needle (BD Bioscience) on the dorsal side of the upper arm just proximal to the elbow. The needle was inserted through the incision/puncture and the virus was injected into the biceps and surrounding connective tissue. For injection into the posterior upper arm, an incision/puncture was made posterior–dorsally just above the elbow, the needle was inserted through the incision/puncture and the virus was injected into the triceps muscles and surrounding connective tissue.

Statistical analysis of outgrowths. First, outgrowths were sorted into several categories according to the extent of outgrowth (see Supplementary Tables 4 and 6). Next, different categories were grouped into two supra-categories: substantial outgrowth versus non-substantial outgrowth, with outgrowths that have at least two cartilaginous segments being the cut-off for what marks a substantial outgrowth. We then used the Cochran–Mantel–Haenszel test with continuity correction for repeated tests of independence on categorical data to assess whether different treatments had statistically significant effects on the likelihood of substantial outgrowths. The two groups of category were: type of manipulation (for example, ABs + FGF8 versus ABs + mCherry) and type of outgrowth (for example, substantial outgrowth versus non-substantial outgrowth).

Choice of sample size in experiments. When we used SAG to investigate the role of HH signalling in accessory limb outgrowth from ABs, the sample size was chosen at random for the first experiment. For subsequent experiments it was chosen based on the rate of accessory limb induction in the first experiment. In experiments involving baculoviral delivery of human *SHH* or axolotl *fgf8* to ABs, the sample size was also chosen based on the rate of accessory limb induction from ABs by SAG.

In experiments testing the potential of SAG to induce accessory limbs from PBs, we used four conditions: PBs treated with SAG, PBs treated with water (negative controls), PBs complemented with anterior skin (positive controls), PBs complemented with posterior skin (negative controls). The sample size for all groups was based on the rate of accessory limb induction from ABs by SAG, as described above.

In experiments involving baculoviral delivery of *mCherry* or *fgf8* to PBs, the sample size was chosen based on the induction rate of accessory limbs from PBs complemented with anterior skin, taking into consideration the false positive rate of accessory limb induction from PBs treated with water, as identified in the PB experiments described above.

In experiments where the determination state of SAG-induced accessory limbs from ABs was assayed by transplanting accessory limb skin onto host anterior or posterior wounds, the sample size was based on preliminary experiments and resulting information about the induction rate of accessory limb outgrowth from PBs complemented with anterior skin or ABs complemented with posterior skin.

Nanostring gene expression analysis. For probe design, sequences of genes of interest were taken from the axolotl transcriptome that has been assembled from Illumina sequencing reads (to be published elsewhere). For genes whose sequence was validated by the Sanger method, the Nanostring probes were designed to the whole validated sequence. For non-validated genes, the Nanostring probes were designed only to the open reading frame of the sequence. Probes were designed by Nanostring Technologies.

RNA isolation for Nanostring analysis was done by a tandem purification using Trizol (Ambion) and Qiagen Micro Kit (Qiagen). The tissue was homogenized in Trizol using a polytron homogenizer. After chloroform addition and centrifugation, the aqueous phase was collected. An equal volume of 70% ethanol was added to the collected aqueous phase and vortexed. The solution was then loaded

© 2016 Macmillan Publishers Limited. All rights reserved

onto a Qiagen Micro Kit column, and RNA purification was done following the manufacturer's protocol.

For Nanostring hybridization, each hybridized sample corresponded to one biological replicate. Each biological replicate consisted of a single blastema, AB or PB. In each hybridization we used 200 ng of RNA (analysis of blastemas) or 300 ng of RNA (analysis of ABs and PBs). Samples were hybridized according to the manufacturer's instructions, and subsequently analysed on the Nanostring Digital Analyzer to obtain the data on the number of counts. Data were normalized based on the counts of housekeeping genes *rpl4* and *ef1* with Nanostring nSolver software. Statistical analysis (one-way ANOVA with false discovery rate correction, followed by Tukey post-hoc test) was performed in Genespring GX (Agilent) software. Graphs showing expression profiles of genes were generated with Prism6 (Graphpad) software.

Alcian blue/alizarin red staining. The regenerates were fixed overnight in MEMFA (1 × MEM salts, 3.7% formaldehyde), followed by three 5-min washes in PBS. The limbs were dehydrated in 25% and 50% ethanol (EtOH) for 10 min each. They were then placed in alcian blue solution and kept at 37 °C for approximately 1 h or until all cartilage elements were visibly stained. The limbs were then washed in EtOH–acetic acid mix for 1 h, 95% EtOH for 15 min and twice in 1% KOH solution for 10 min each time, until the soft tissue became clear. They were then placed in alizarin red solution, and stained at room temperature until the ossified parts of the bones became visibly stained (usually approximately 25 min). They were subsequently washed in 1% KOH–glycerol mix. Finally, the limbs were placed in 20% glycerol–PBS solution and imaged using an Olympus SZX16 microscope and Cell F software (Olympus).

Cryosectioning. Limbs with accessory limbs were collected and fixed in MEMFA overnight at 4 °C. Afterwards, the limbs were washed for 2 × 5 min and 4 × 10 min in 1× PBS and placed in 20% sucrose solution overnight at 4 °C. The next day they were embedded in optimum cutting temperature (OCT) compound and frozen in plastic blocks. Samples were sectioned into 10-μm-thick slices that were collected on Superfrost-plus microscopy slides (Menzel-Gläser). The sections were allowed to air dry for 1 h before being stored at −20 °C.

Immunostaining. The following antibodies were used for staining: anti-Meis1/2/3, clone 9.2.7 (EMD Millipore); anti-beta-3-tubulin (R&D systems MAB1195); anti-myosin heavy chain (generated in-house at MPI-CBG antibody facility). Staining was performed using protocols previously described[31,32].

Paraffin embedding and sectioning. Limbs with accessory limbs or blastemas were collected and fixed in MEMFA for 3 days at 4 °C. Afterwards, the limbs were washed for 2 × 5 min in 1× PBS treated with diethylpyrocarbonate (DEPC) and sequentially dehydrated in 25%, 50%, 75% and 100% ethanol in DEPC-treated water. The samples were left overnight in 100% ethanol. The next day, the limbs were washed for 10 min in fresh 100% ethanol, then for 3× 40 min to1 h in 100% xylol at room temperature, and lastly for 3 × 1 h in melted paraffin at 70 °C. The limbs were placed on the bottom of disposable plastic base moulds filled with molten paraffin and left on a smooth surface to solidify at room temperature. Samples were cut into sections with a thickness of 7–10 μm and transferred onto superfrost-plus microscopy slides (Menzel-Gläser) covered with DEPC-treated water. The slides were placed on a heating platform at 37 °C and allowed to air dry for at least 1 h.

Isolation of genes for *in situ* hybridization. We used 10–12 freshly frozen 9-day or 15-day blastemas for total RNA extraction using a standard trizol-based protocol. A SuperScript III kit (Invitrogen) was used for cDNA synthesis according the manufacturer's protocol. cDNA was used for PCR with appropriate primers to isolate the genes of interest. PCR products were run in 2% agarose gel and bands of the expected size were cut out and purified from the gel. PCR products were cloned into a PCR-II Topo vector, using TOPO TA Cloning kit (Invitrogen). Accession numbers are listed after the Acknowledgements section.

For *grem1 in situ* hybridization we pulled out a sequence of 937 bp by PCR using the following primers: forward TCGCCTGACACTACATAGC, reverse AGATGTGCAAAAGTTCAGAGAT. For *hand2 in situ* hybridization we pulled out a sequence of 632 bp by PCR using the following primers based on the newt HAND2 sequence: forward TGAGCCTGGTGGGGGGCTTCC, reverse AGCGCCCAGACGTGCTGCGGCCA. For *gli3 in situ* hybridization we isolated a sequence of 933 bp by PCR using the following primers: forward ATCATTAATAAAGAAGGAGATCCA, reverse ATAGGTCTCTGGGTAGGAAAAG. For *hoxd13 in situ* hybridization we isolated a sequence of 735 bp by PCR using the following primers: forward GGCCGAGGTTAGTTTTTATAACCGCA, reverse AAAGGAAGACTTCCAGAAGTGCGAGCTCT. For *fgf8 in situ* hybridization we isolated a sequence of 1,148 bp by PCR using the following primers: forward TTCACGTGCCTCACCTCCACTACCTCAGCA, reverse ATATAAATTGTTCTTCTAAAAGTCCACTGG. For *fgf9 in situ* hybridization we isolated a sequence of 972 bp by PCR using the following primers: forward ATTAAACCCGCTTTGATTTCTTG, reverse GTAGAGATTAGATG AGTATGTGTTA. *Fgf10* and *fgf17* sequences were isolated by screening an EST library of 6-day axolotl tail blastema cDNA. For *ptc1* and *shh in situ* hybridization, we used sequences previously described[33].

***In situ* hybridization probe synthesis.** The probes were synthesized using linearized plasmids containing genes of interest, SP6 or T7 polymerase (produced in-house), and DIG-labelled nucleotide mixture (Roche). Synthesis was performed for 2–5 h at 37 °C. RNA was purified from the RNA synthesis mix using RNeasy MinElute Cleanup Kit (Qiagen) and a small amount of purified RNA was run on a 1% agarose gel to assess its quality. Concentration was measured on NanoDrop and adjusted to 100 ng μl^{-1} with hybridization buffer (50% formamide, 10% dextran, 5× SSC, 0.1% Tween, 1 mg ml^{-1} yeast RNA, 100 mg ml^{-1} heparin, 1× Denhardt's solution, 0.1% CHAPS detergent and 5 mM EDTA).

***In situ* hybridization.** Slides were deparaffinized by washing 3 times in 100% xylol and 3 times in 100% ethanol for 10 min each step, dried for 15 min and placed overnight at 70 °C in hybridization buffer containing probes at the following concentrations: *fgf8* 500 ng μl^{-1}, *fgf9* 300 ng μl^{-1}, *fgf10* 300 ng μl^{-1}, *fgf17* 300 ng μl^{-1}, *gli3* 500 ng μl^{-1}, *grem1* 500 ng μl^{-1}, *hand2* 200 ng μl^{-1}, *shh* 200 ng μl^{-1}. On the next day, slides were washed in salt solutions at 70 °C as follows: 30 min in 5 × SSC, 40 min in 2 × SSC twice, 20 min in 0.2 × SSC. To reduce background signal coming from nonspecific probe binding, slides were washed in Tris–sodium–EDTA (TNE) solution for 2 × 10 min, treated with 20 μg ml^{-1} RNase A (Sigma) in TNE for 15 min, and washed in TNE twice, for 2 min and 10 min. After RNase treatment, slides were washed for 2 × 5 min and 1 × 10 min in maleic acid buffer (MAB) and blocked for 1 h with blocking buffer (1% blocking reagent (Roche) dissolved in MAB). Then slides were incubated with anti-DIG-AP antibody (Roche) diluted 1:5,000 in blocking buffer for 1–2 h at room temperature or overnight at 4 °C. Slides were washed at room temperature in MAB for 2 min, 5 min and 3 × 10 min, followed by 2 × 10 min washes in AP buffer. Slides were developed in BM Purple solution (Roche) or preferentially a substrate working solution containing NBT/BCIP (Roche) and 5% polyvinyl alcohol (Sigma) at 37 °C in the dark until the signal became clearly visible. The reaction was stopped by washing in 1 × PBS several times.

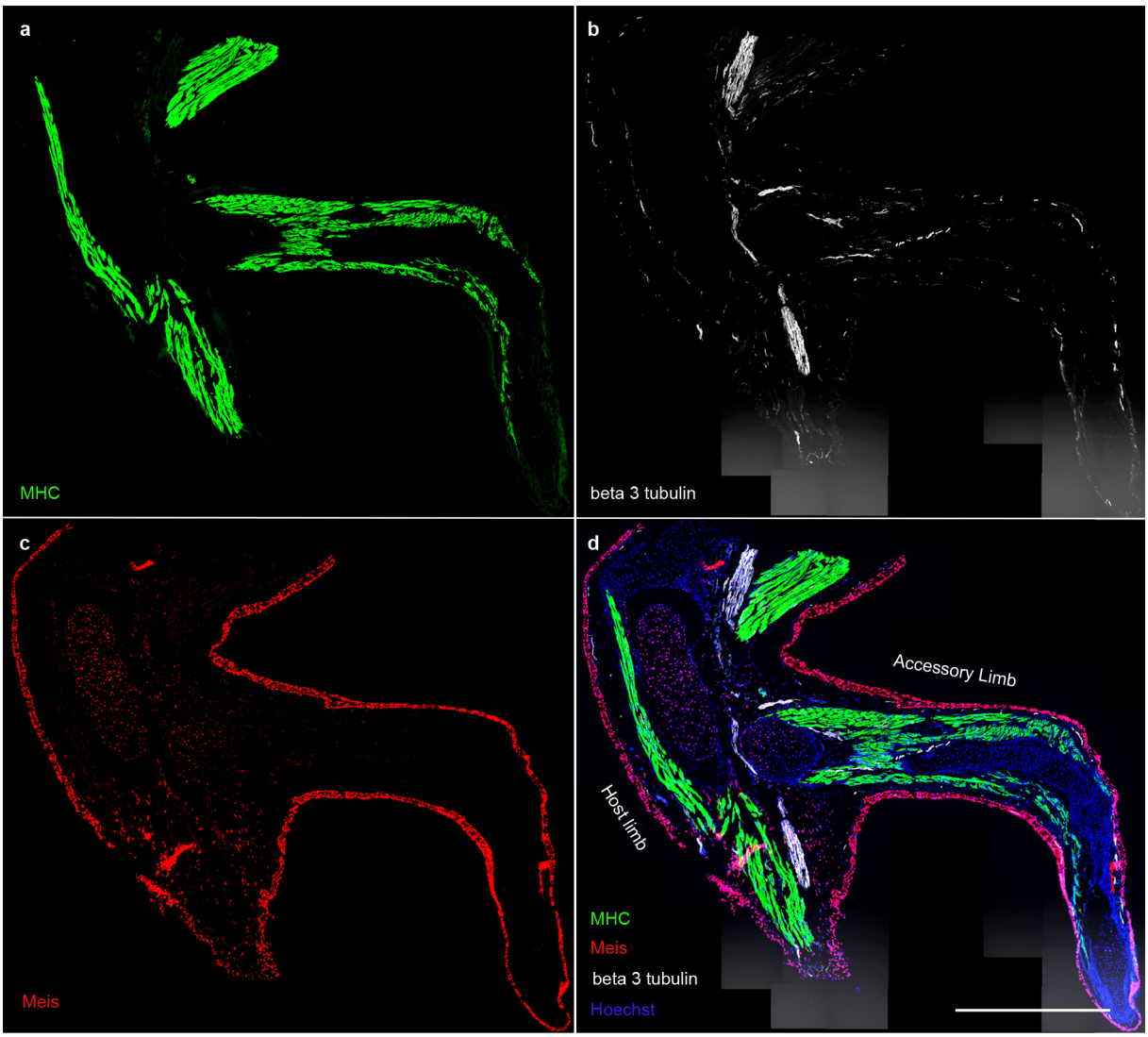
In the case of the *fgf9 in situ* hybridization signal in untreated 12-day ABs (Extended Data Fig. 4f), the signal was localized to the base of the blastema but was not considerably stronger than the background signal that is typically seen in muscle and nerve for limb *in situ* hybridization samples. Combined with observed reduction in counts of *fgf9* after 8 days in Nanostring data (Fig. 1d), we cannot conclusively say whether this basal *fgf9* signal represents *bona fide* expression or background signal.

Imaging and figure assembly. Images of antibody-stained samples were acquired on a Zeiss Axioobserver Z.1 using AxioVision 4.8.1.0 software (Zeiss). Images of *in situ*-stained samples were acquired on an Olympus BX61VS110 system using Olympus VS-ASW software.

Images were processed with Fiji or Photoshop software. When adjusted, contrast and brightness were adjusted uniformly across the image.

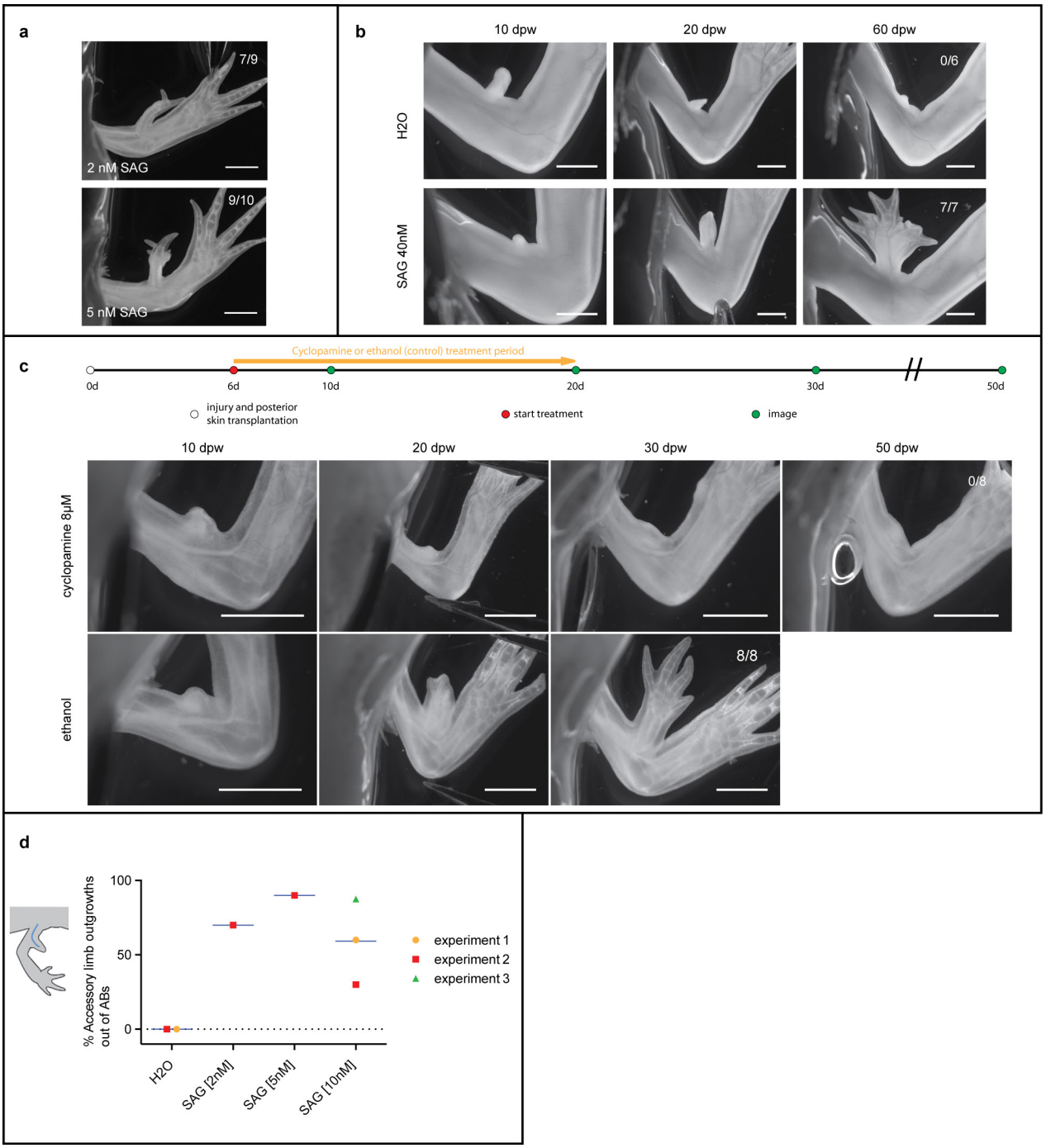
26. Sobkow, L., Epperlein, H.-H., Herklotz, S., Straube, W. L. & Tanaka, E. M. A germline GFP transgenic axolotl and its use to track cell fate: dual origin of the fin mesenchyme during development and the fate of blood cells during regeneration. *Dev. Biol.* **290,** 386–397 (2006).
27. Boyce, F. M. & Bucher, N. L. Baculovirus-mediated gene transfer into mammalian cells. *Proc. Natl Acad. Sci. USA* **93,** 2348–2352 (1996).
28. Kost, T. A. & Condreay, J. P. Recombinant baculoviruses as mammalian cell gene-delivery vectors. *Trends Biotechnol.* **20,** 173–180 (2002).
29. Kaikkonen, M. U. *et al.* Truncated vesicular stomatitis virus G protein improves baculovirus transduction efficiency *in vitro* and *in vivo*. *Gene Ther.* **13,** 304–312 (2006).
30. Hopkins, R. & Esposito, D. A rapid method for titrating baculovirus stocks using the Sf-9 Easy Titer cell line. *BioTechniques* **47,** 785–788 (2009).
31. Kragl, M. *et al.* Cells keep a memory of their tissue origin during axolotl limb regeneration. *Nature* **460,** 60–65 (2009).
32. Nacu, E. *et al.* Connective tissue cells, but not muscle cells, are involved in establishing the proximo-distal outcome of limb regeneration in the axolotl. *Development* **140,** 513–518 (2013).
33. Schnapp, E., Kragl, M., Rubin, L. & Tanaka, E. M. Hedgehog signaling controls dorsoventral patterning, blastema cell proliferation and cartilage induction during axolotl tail regeneration. *Development* **132,** 3243–3253 (2005).

© 2016 Macmillan Publishers Limited. All rights reserved



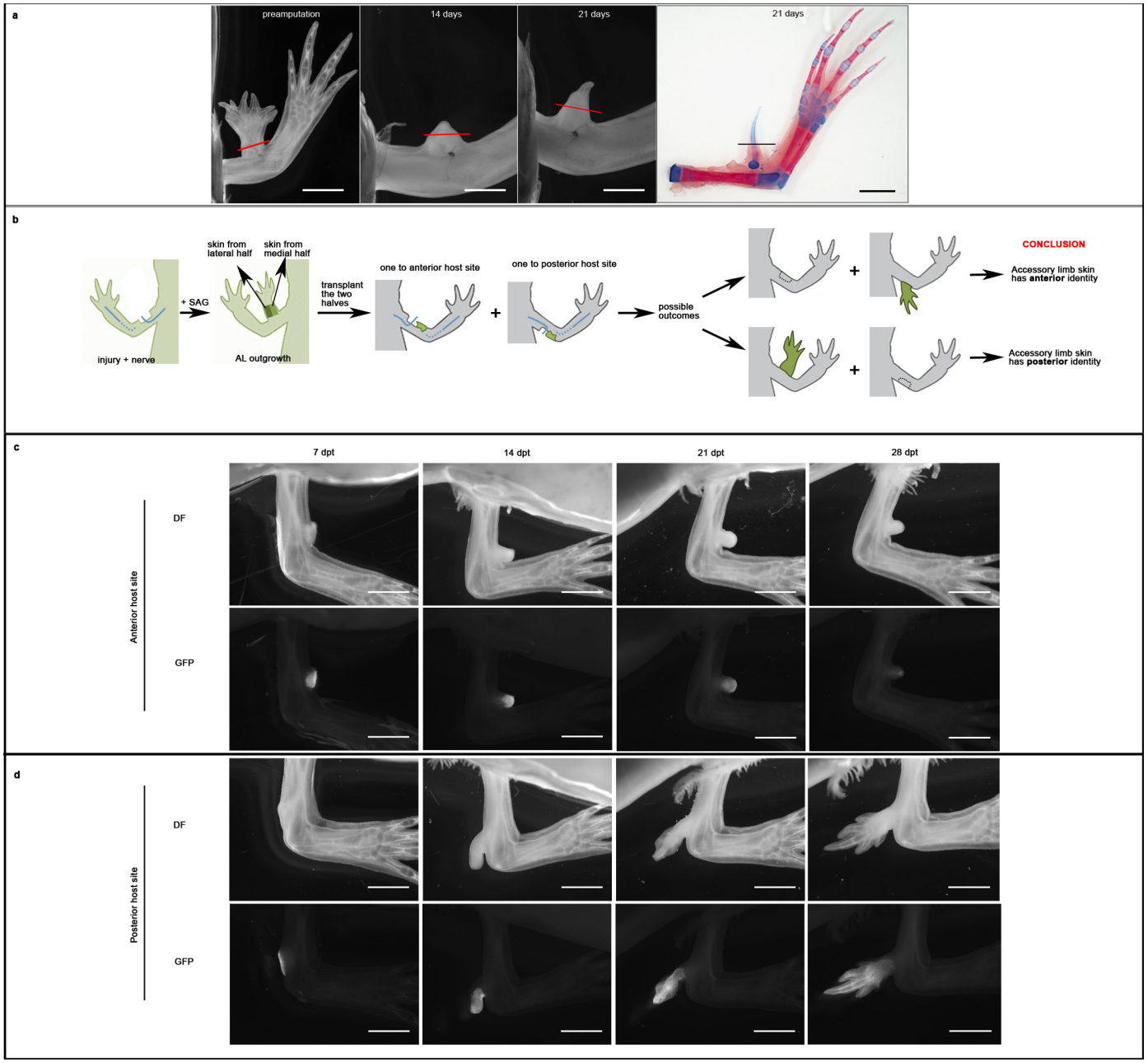
Extended Data Figure 1 | SAG-induced accessory limbs contain muscles, axons, and primarily lower arm and hand elements. **a**, Immunostaining for myosin heavy chain (MHC) reveals muscle fibres in the accessory limb (AL). **b**, Immunostaining for β3-tubulin reveals axons in the accessory limb. **c**, Immunostaining for MEIS, an upper arm identity marker, reveals that the accessory limb contains primarily lower arm and hand elements, as evidenced by the absence of MEIS in the accessory limb. Expression of MEIS is seen in the host upper arm on which the accessory limb was induced. MEIS staining is also observed in the epidermis, however epidermis does not play a role in positional identity. **d**, Overlay of anti-MHC (green), anti-β3-tubulin (white), anti-MEIS (red) and Hoechst (marks nuclei, blue) staining. Number of accessory limbs analysed, $n = 5$. Scale bar, 1 mm.

© 2016 Macmillan Publishers Limited. All rights reserved



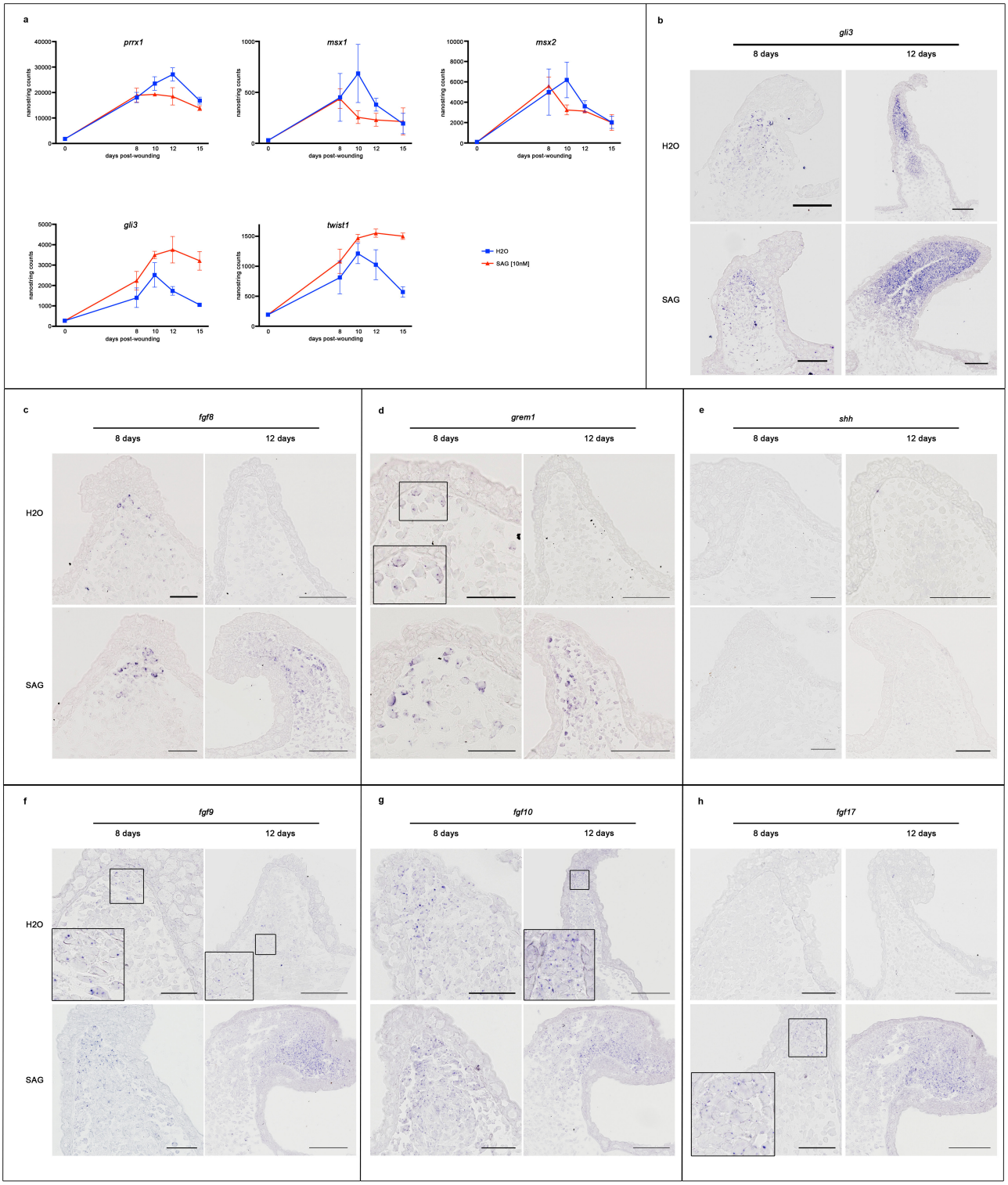
Extended Data Figure 2 | HH signalling is sufficient for accessory limb outgrowth from ABs and is necessary for outgrowth from ABs complemented with posterior skin. a, Treatment of ABs with lower concentrations of SAG still results in accessory limb outgrowth, but with a decreased number of digits. Data from a single experiment. **b**, SAG (40 nM) also induces accessory limb outgrowth from ABs in large animals (13–14 cm snout to tail tip), indicating that this effect is not size-dependent. Data from a single experiment. **c**, Blocking HH signalling with cyclopamine (8 µM) in ABs complemented with posterior skin blocks regeneration. Data from a single experiment. **d**, Percentage of accessory limb outgrowths from ABs treated with different concentrations of SAG. See Supplementary Table 1 for details. Short blue horizontal lines represent average percentages for each condition. The dotted line is for visual aid and intersects the *y* axis at 0%. Numbers in upper right corners of photomicrographs indicate the number of accessory limb outgrowths out of the total number of ABs. dpw, days post-wounding. Scale bars, 2 mm.

© 2016 Macmillan Publishers Limited. All rights reserved



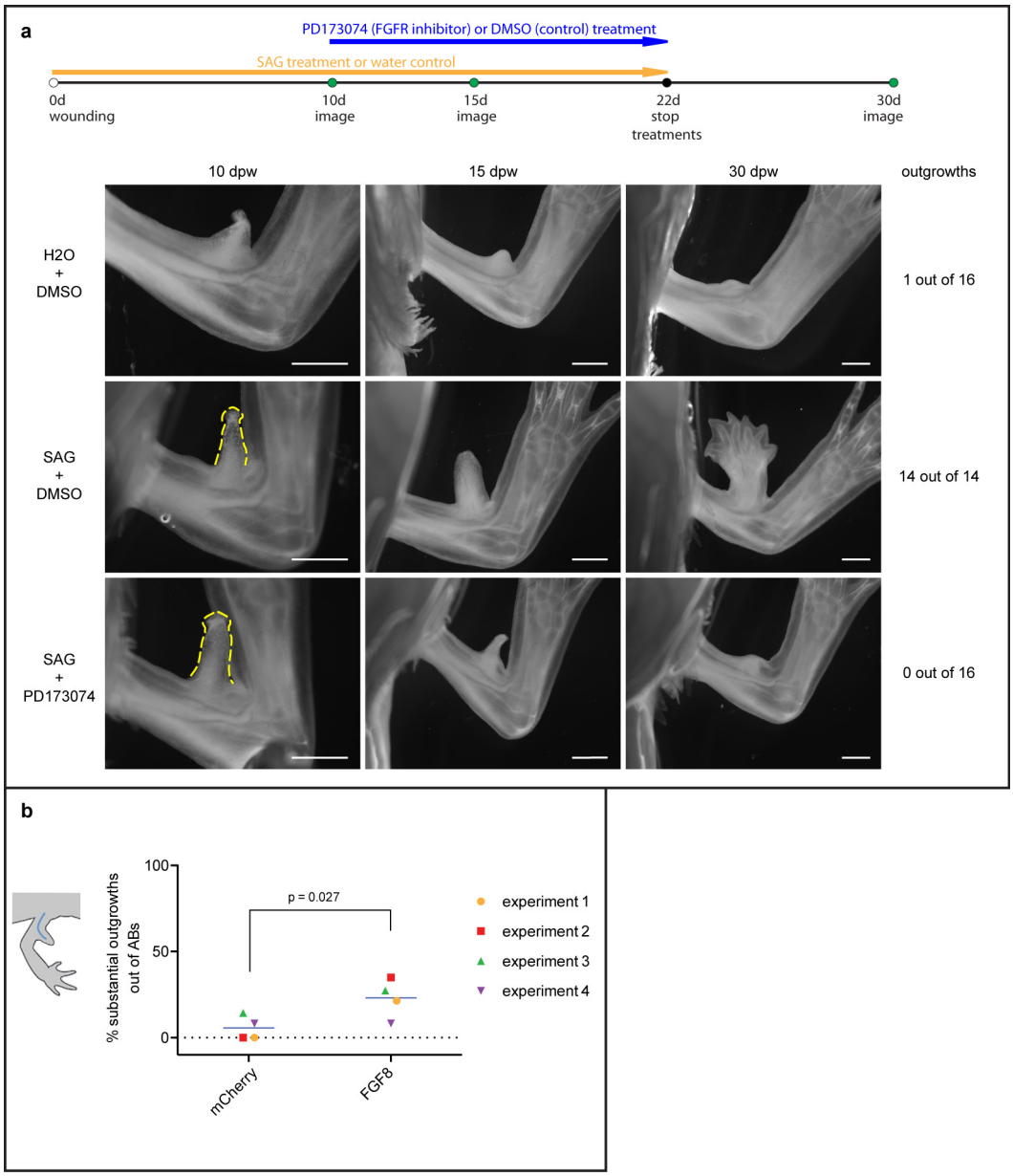
Extended Data Figure 3 | SAG treatment does not make the determination state of accessory limbs more posterior. a, SAG-induced accessory limbs (SAG-AL) regenerate at most a spike after amputation, indicating a lack of functional positional discontinuity. SAG-ALs were amputated and imaged at 14 and 21 days post-amputation. No further growth of the spike was observed 21 days post-amputation. Red or black lines, amputation plane. No SAG-ALs regenerated a patterned accessory limb after amputation ($n = 0$ of 5). Experimental data from a single experiment. Alcian blue/alizarin red staining: cartilage is blue, ossifications are red. **b**–**d**, Testing determination state of SAG-ALs by skin transplantation to ABs or PBs. **b**, Schema illustrating the experimental setup and potential outcomes. Accessory limbs were generated by SAG treatment of ABs from transgenic animals constitutively expressing GFP. The SAG-ALs were divided into a medial and lateral side, with skin from one side transplanted to an anterior host site and skin from the opposite side transplanted to a posterior host site. Two possible outcomes were expected, depending on whether SAG posteriorized cells during initial accessory limb outgrowth. **c**, A host AB with a skin transplant from a SAG-AL shows no accessory limb outgrowth, indicating no determination of posterior identity during SAG-AL outgrowth ($n = 0$ of 22). **d**, A host PB with a skin transplant from the same SAG-AL as in **c** results in new accessory limb outgrowth, revealing the anterior identity of the transplanted skin ($n = 15$ of 21). Data from a single experiment. See Supplementary Table 2 for details. DF, dark field images. GFP panels show the transplanted skin. dpt, days post transplantation of skin. Scale bars, 2 mm.

© 2016 Macmillan Publishers Limited. All rights reserved



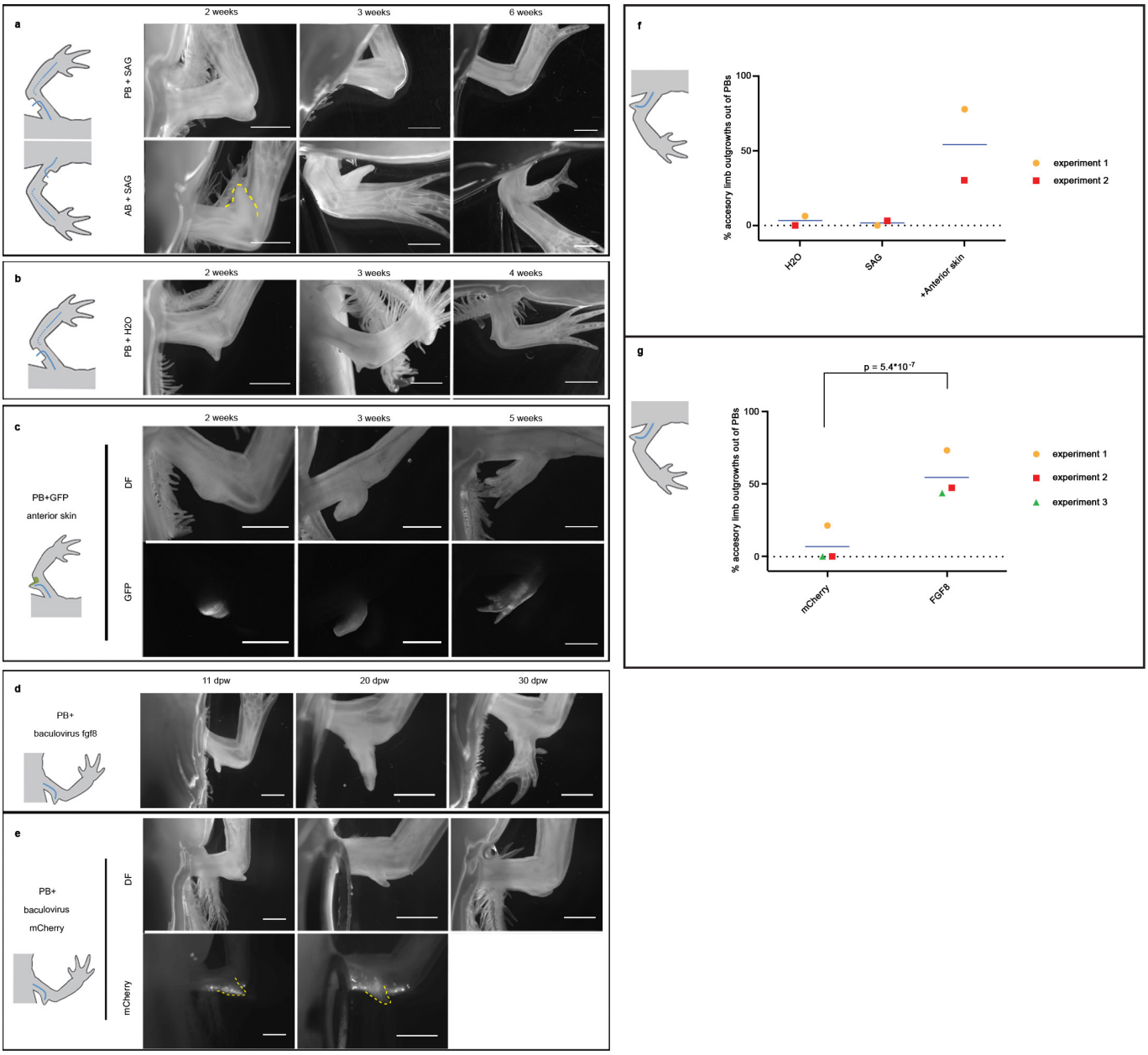
Extended Data Figure 4 | Gene expression analysis of ABs with and without SAG. a, ABs express blastema markers *prrx1*, *msx1*, *msx2*, and *twist1*, and anterior marker *gli3*. The *y* axis indicates the number of normalized counts obtained by Nanostring nCounter technology. Each data point is the mean normalized count number from biological replicates ($n = 3$ or 4; for individual n values and complete data set, see Supplementary Table 7a; data from one experiment). Error bars represent s.d. For data points where s.d. was small, no error bars are shown. **b**, *gli3* is expressed at 8 and 12 days in SAG-treated or untreated ABs. **c**–**h**, Expression of FGF–SHH loop components in SAG-treated versus untreated ABs. **c**, **d**, **f**, In ABs, *fgf8*, *fgf9*, and *grem1* maintenance, but not induction, requires SAG treatment. **c**, **d**, *fgf8* and *grem1* are upregulated in ABs in the presence or absence of SAG at 8 days, but are downregulated at 12 days unless SAG is present. **f**, *fgf9* is expressed at the tip of the bumps at 8 days and maintains this expression at the tip in 12-day SAG-treated samples. In untreated 12-day samples there is no more *fgf9* expression at the tip, although there seems to be expression at the base of the bump. This could represent a secondary source of *fgf9* that is not part of the FGF–SHH circuitry or it could be an unspecific signal; see Methods. **g**, *fgf10* was expressed in all conditions. **h**, Unlike other *fgf* genes, *fgf17* expression required SAG at all time points. *fgf17* was not expressed in untreated samples at either 8 or 12 days, but was expressed at both time points when treated with SAG. **e**, SAG is not sufficient to upregulate *shh* in ABs. *Shh* is not expressed in ABs at 8 or 12 days in the presence or absence of SAG. For detailed information on the number of samples analysed and results, see Supplementary Table 3. Inserts show magnification of positive cells. Images of 12-day SAG-treated blastemas in panels **f**–**h** show different sections from the same limb that were hybridized separately for each of the indicated genes. Scale bars: all 8-day images, 100 μm; all 12-day images, 200 μm.

© 2016 Macmillan Publishers Limited. All rights reserved



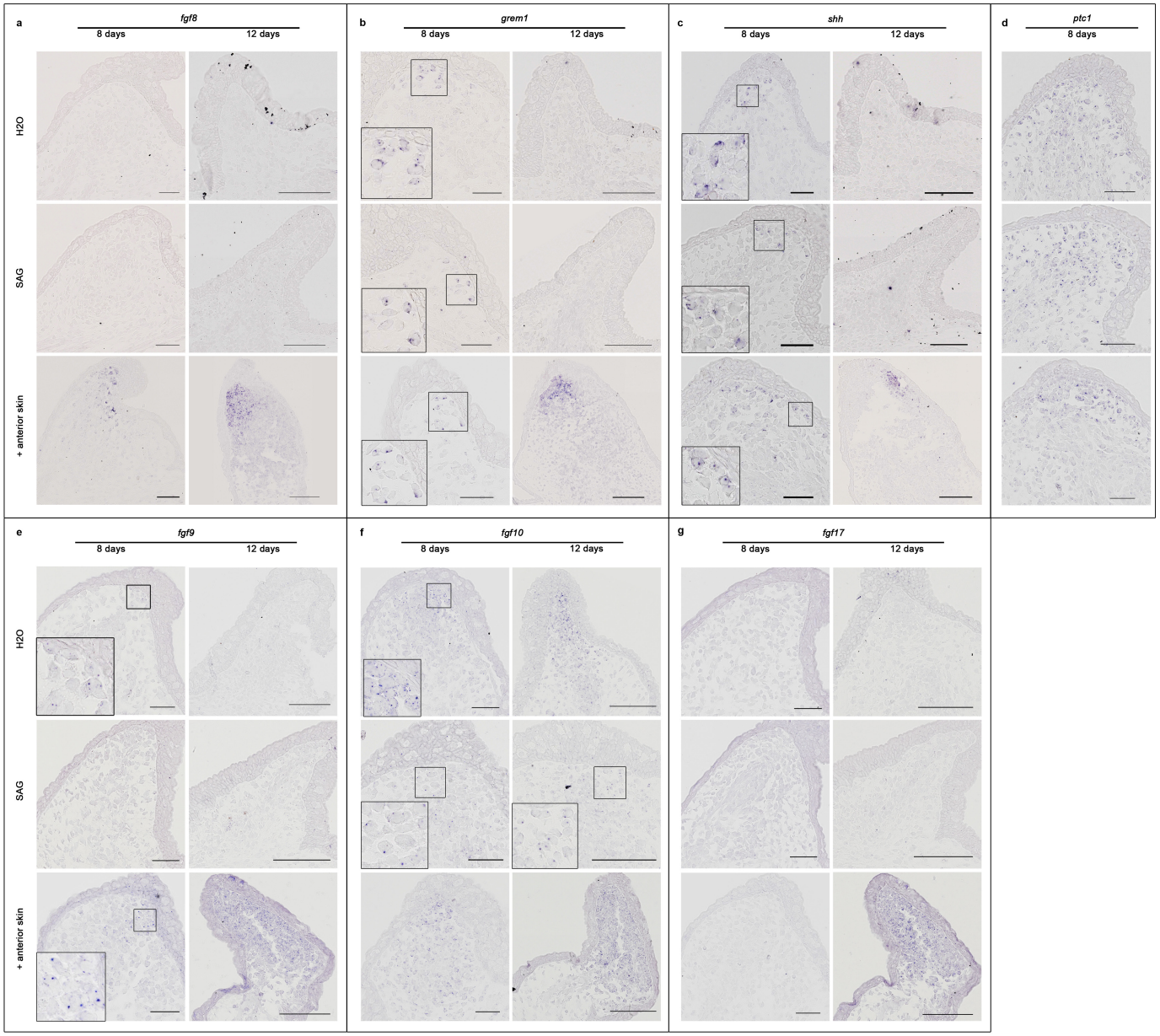
Extended Data Figure 5 | FGF signalling is required for SAG-induced accessory limb outgrowth from ABs, and is sufficient to induce single-digit accessory limb outgrowth from ABs. a, FGF signalling is required for SAG-induced accessory limb outgrowth. ABs were treated with 15 nM SAG or water starting on the day of wounding. At 10 days post-wounding, concomitant treatment with DMSO or 15 μM PD173074 (inhibitor of FGFR1 and FGFR3) was initiated. Treatment was terminated at 22 days post-wounding. Images of ABs were taken at 10, 15 and 30 days post-wounding. ABs treated with water and DMSO fail to develop accessory limbs ($n = 1$ of 16). ABs treated with SAG and DMSO form accessory limbs ($n = 14$ of 14). ABs treated with SAG and PD173074 grow until initiation of PD173074 treatment, after which growth halts and ABs fail to develop accessory limbs ($n = 0$ of 16), eventually regressing. This indicates that FGF signalling acts downstream of SAG and is required for accessory limb outgrowth. Dashed yellow lines demarcate ABs. The experimental data come from two experiments. dpw, days post-wounding. Scale bars, 1 mm. **b**, Ectopic baculovirus-induced expression of FGF8 in ABs results in single digit accessory limb outgrowths. Percentage of substantial outgrowths from ABs transduced with *fgf8* or *mCherry* (negative control). See Supplementary Table 4 for details. Short blue horizontal lines represent average percentage for each condition. The dotted line is for visual aid and intersects the *y* axis at of 0%.

© 2016 Macmillan Publishers Limited. All rights reserved



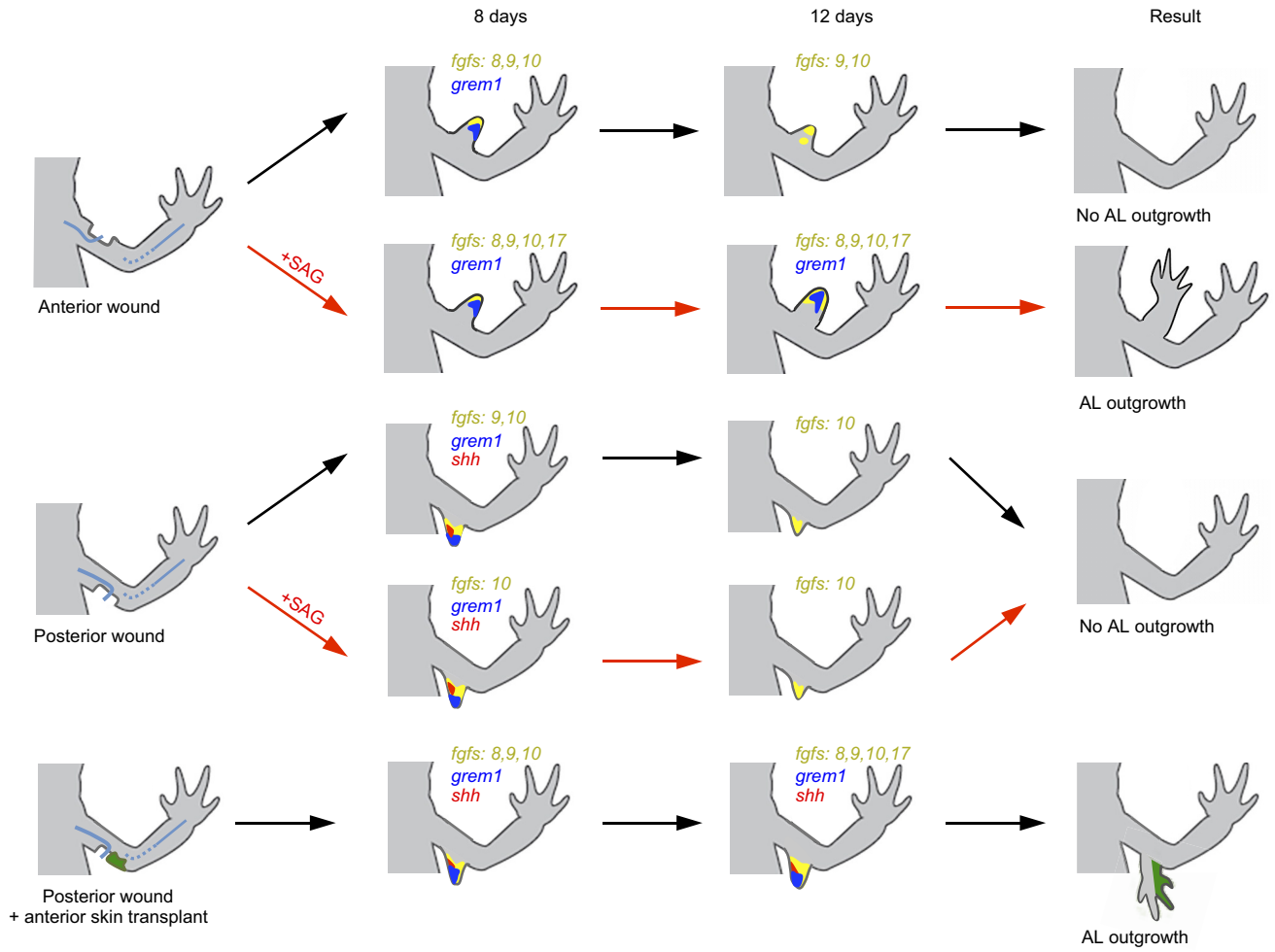
Extended Data Figure 6 | FGF8, but not HH signalling, is sufficient to drive accessory limb outgrowth from PBs. a–**c**, **f**, HH signalling is not sufficient to drive accessory limb outgrowth from PBs (for details see Supplementary Table 1). **a**, SAG is sufficient to drive accessory limb outgrowth from ABs ($n = 23$ of 32) but not PBs ($n = 1$ of 32) on the contralateral limb of the same animal. Data from one experiment. The images are the same as in Fig. 2a. **b**, Untreated PBs also produce no accessory limb outgrowth ($n = 1$ of 37; data from two experiments). **c**, Positive control: PBs complemented with anterior skin transplants from animals that constitutively express GFP grew into accessory limbs without SAG treatment ($n = 10$ of 19, data from two experiments). **f**, Percentage of accessory limb outgrowths from PBs without treatment (H_2O), treated with SAG, or complemented with anterior skin. **d**, **e**, **g**, FGF8 is sufficient to drive accessory limb outgrowth from PBs (data from three experiments; for details see Supplementary Table 5). **d**, PBs expressing baculovirus-induced FGF8 grow into accessory limbs ($n = 27$ of 50). The images are the same as in Fig. 2b. **e**, Negative control: PBs expressing baculovirus-induced mCherry do not grow into accessory limbs ($n = 3$ of 49). **g**, Percentage of accessory limb outgrowths from PBs transduced with *fgf8* or *mCherry* (negative control). Short blue horizontal lines in all graphs represent average percentage for each condition. Dotted lines are for visual aid and intersect the *y* axes at 0%. Dashed yellow lines demarcate blastemas. DF, dark field images; dpw, days post-wounding. Scale bars, 2 mm.

© 2016 Macmillan Publishers Limited. All rights reserved



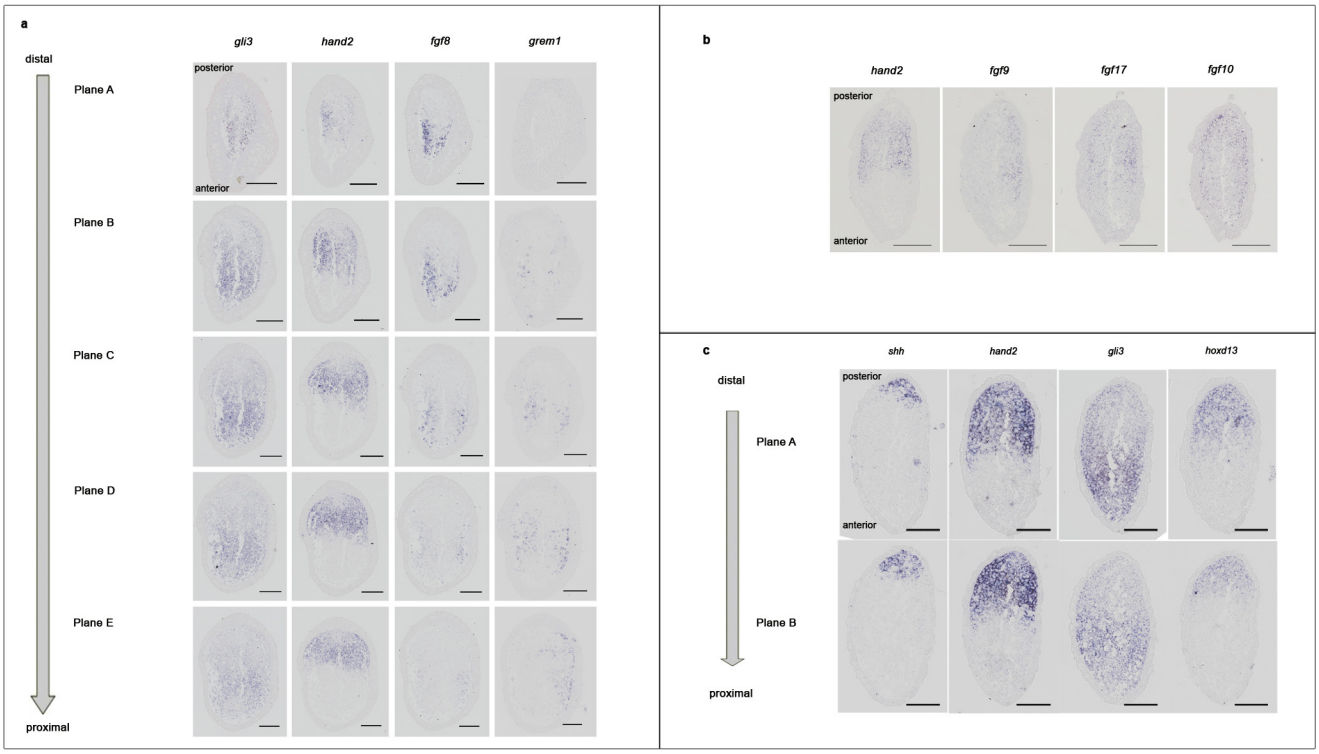
Extended Data Figure 7 | PBs do not express *fgf8* or *fgf17* even though intact HH signalling is present, but do express *grem1*, *fgf10* and *fgf9*. **a**, **g**, PBs do not express *fgf8* and *fgf17* even upon SAG treatment. *Fgf8* and *fgf17* are not expressed in PBs at 8 or 12 days in the presence or absence of SAG, but are expressed in positive control PBs containing an anterior skin transplant. **b**–**d**, **f**, Early induction of *grem1*, *shh*, *ptc1* (a HH signalling target) and *fgf10* in posterior wounds is independent of SAG, and SAG is not sufficient for maintenance of *grem1* and *shh* expression. **b**, **c**, *Grem1* and *shh* are present in PBs at 8 days and are downregulated by 12 days independent of SAG treatment. This observation differs slightly from the Nanostring nCounter data (Fig. 2d) where *grem1* is downregulated at 15 days. The discrepancy is probably due to the variation in regeneration timelines across different experiments. *Grem1* and *shh* are present in control PBs containing an anterior skin transplant at both 8 and 12 days. **d**, *Ptc1* is expressed at 8 days in PBs that were untreated or treated with SAG or complemented with an anterior skin graft. **f**, *Fgf10* is present in all conditions at 8 and 12 days. **e**, *Fgf9* is expressed in a few cells at the tip of 8-day untreated PBs, but is absent from 12-day untreated PBs and 8- or 12-day SAG-treated PBs. *Fgf9* is present in control PBs complemented with anterior skin grafts. Detailed information on the number of samples analysed and results can be found in Supplementary Table 3. Inserts show magnification of positive cells. Images of 12-day PBs complemented with anterior skin (**e**–**g**) show different sections from the same limb that were hybridized separately for each of the indicated genes. Scale bars: all 8-day samples, 100 μm; all 12-day samples, 200 μm.

© 2016 Macmillan Publishers Limited. All rights reserved



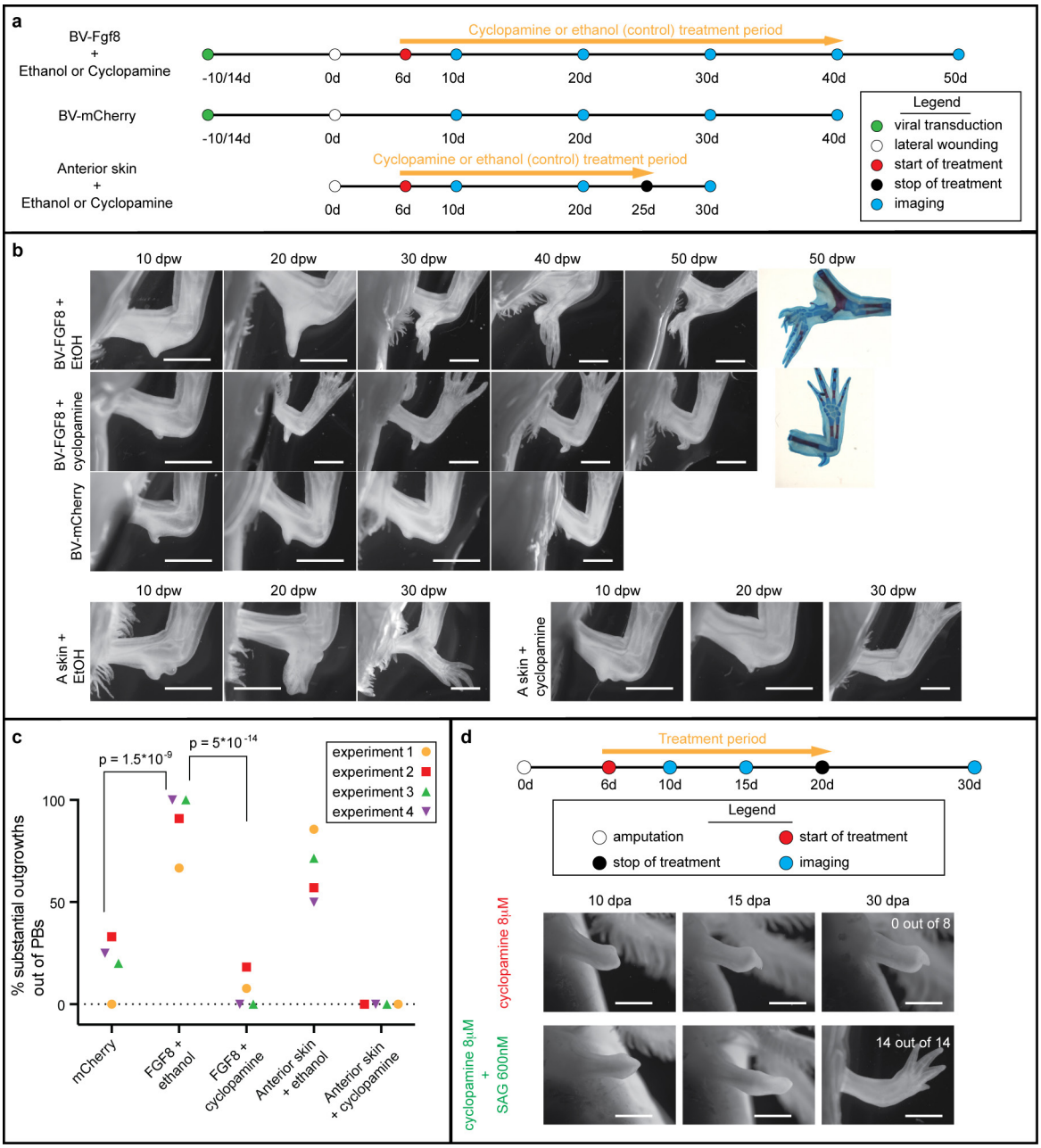
Extended Data Figure 8 | Schematic illustration of the expression pattern of *fgf8*, *fgf9*, *fgf10*, *fgf17*, *grem1* and *shh* in ABs and PBs treated with SAG or untreated and PBs complemented with anterior skin transplant. ABs upregulate *fgf8*, *fgf9*, *fgf10* and *grem1* but not *shh* in the absence of SAG at 8 days post-wounding, indicating that initial upregulation of *fgf8*, *fgf9* and *fgf10* (yellow) and *grem1* (blue) is independent of SAG. ABs downregulate expression of *fgf8* and *grem1* by 12 days, and eventually regress. ABs treated with SAG maintain *fgf8* and *grem1* expression at 12 days, indicating that HH signalling is necessary for their maintenance. SAG-treated ABs also express *fgf17*. ABs treated with SAG eventually develop accessory limbs (ALs). Untreated PBs upregulate *fgf9*, *fgf10*, *grem1* and *shh* (red) but not *fgf8* nor *fgf17*. SAG-treated PBs also do not upregulate *fgf8* or *fgf17* and do not produce accessory limb outgrowths, indicating that SHH signalling is not sufficient for accessory limb outgrowth from PBs. PBs complemented with an anterior skin transplant express *fgf8*, *fgf9*, *fgf10*, *fgf17*, *grem1* and *shh* and form accessory limbs.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 9 | Expression patterns of *gli3*, *hand2*, *fgf8*, *fgf9*, *fgf10*, *fgf17*, *grem1*, *shh* and *hoxd13* in normal limb regeneration. **a**, Expression of *Gli3*, *hand2*, *fgf8* and *grem1* was analysed on adjacent sections of a 10-day blastema. Sections are arranged from the distal tip of the blastema (top, plane A) to the proximal base of the blastema (bottom, plane E). *Gli3* is expressed throughout the blastema, with stronger expression in the anterior half that lacks *hand2* expression and weaker expression in the *hand2*-expressing zone. *Fgf8* is expressed in the anterior half that lacks *hand2* expression. *Grem1* is expressed more proximally than the other genes (starting from plane B). It is primarily expressed in the anterior half of the blastema, but is also expressed in the posterior half, particularly in the proximal part (planes D and E). **b**, Unlike *fgf8*; *fgf9*, *fgf10* and *fgf17* are not excluded from the *hand2*-expressing, posterior half of a blastema (11 days). **c**, *Shh* is expressed in a small posterior domain in the blastema (12 days). The planes are sections of the same blastema along the proximo-distal axis. Scale bars, 200 μm.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 10 | HH signalling is required for FGF8-induced accessory limb outgrowth from PBs. a, Schematic outline of the experiment. Each replicate experiment consisted of transducing limbs with *fgf8* or *mCherry* (negative control), followed by injury with nerve deviation at 10 or 14 days after transduction. Treatment of FGF8-expressing animals with 8 μM cyclopamine or an equivalent amount of ethanol (cyclopamine solvent) was started at 6 days post-wounding. Additionally a group of non-transduced PBs were complemented with anterior skin and treated with 8 μM cyclopamine (as controls for blocking HH signalling) or ethanol. **b**, Examples of typical outcomes in each condition. Images from the condition PB+FGF8+cyclopamine are the same as in Fig. 2c. Scale bars, 2 mm. dpw, days post-wounding. **c**, Percentage of substantial outgrowths in each condition. Data from four experiments. For detailed description of the number of limbs in each condition and the way they were scored see Supplementary Table 6. The dotted line is for visual aid and intersects the *y* axis at 0%. **d**, The effect of cyclopamine is specific to HH signalling. Limbs were amputated and treatment was started 6 days post-amputation with 8 μM cyclopamine alone or in combination with 600 nM SAG. Treatment with cyclopamine blocked regeneration, which was rescued by co-treatment with SAG. Numbers indicate regenerated limbs out of total amputated limbs. Data from two experiments. dpa, days post-amputation.

© 2016 Macmillan Publishers Limited. All rights reserved

LETTER

doi:10.1038/nature17662

An obligatory role for neurotensin in high-fat-diet-induced obesity

Jing Li[1,2], Jun Song[1,2]*, Yekaterina Y. Zaytseva[2,3]*, Yajuan Liu[2], Piotr Rychahou[1,2], Kai Jiang[2], Marlene E. Starr[1], Ji Tae Kim[1,2], Jennifer W. Harris[1,2], Frederique B. Yiannikouris[4], Wendy S. Katz[4], Peter M. Nilsson[5,6], Marju Orho-Melander[5], Jing Chen[7,8], Haining Zhu[7,8], Timothy Fahrenholz[2,3,9], Richard M. Higashi[2,3,9], Tianyan Gao[2,7], Andrew J. Morris[10], Lisa A. Cassis[4], Teresa W.-M. Fan[2,3,9], Heidi L. Weiss[2,11], Paul R. Dobner[12], Olle Melander[5,6], Jianhang Jia[2,7]§ & B. Mark Evers[1,2]§

Obesity and its associated comorbidities (for example, diabetes mellitus and hepatic steatosis) contribute to approximately 2.5 million deaths annually[1] and are among the most prevalent and challenging conditions confronting the medical profession[2,3]. Neurotensin (NT; also known as NTS), a 13-amino-acid peptide predominantly localized in specialized enteroendocrine cells of the small intestine[4] and released by fat ingestion[5], facilitates fatty acid translocation in rat intestine[6], and stimulates the growth of various cancers[7]. The effects of NT are mediated through three known NT receptors (NTR1, 2 and 3; also known as NTSR1, 2, and NTSR3, respectively)[8]. Increased fasting plasma levels of pro-NT (a stable NT precursor fragment produced in equimolar amounts relative to NT) are associated with increased risk of diabetes, cardiovascular disease and mortality[9]; however, a role for NT as a causative factor in these diseases is unknown. Here we show that NT-deficient mice demonstrate significantly reduced intestinal fat absorption and are protected from obesity, hepatic steatosis and insulin resistance associated with high fat consumption. We further demonstrate that NT attenuates the activation of AMP-activated protein kinase (AMPK) and stimulates fatty acid absorption in mice and in cultured intestinal cells, and that this occurs through a mechanism involving NTR1 and NTR3 (also known as sortilin). Consistent with the findings in mice, expression of NT in *Drosophila* midgut enteroendocrine cells results in increased lipid accumulation in the midgut, fat body, and oenocytes (specialized hepatocyte-like cells) and decreased AMPK activation. Remarkably, in humans, we show that both obese and insulin-resistant subjects have elevated plasma concentrations of pro-NT, and in longitudinal studies among non-obese subjects, high levels of pro-NT denote a doubling of the risk of developing obesity later in life. Our findings directly link NT with increased fat absorption and obesity and suggest that NT may provide a prognostic marker of future obesity and a potential target for prevention and treatment.

To examine the potential role of NT in fat deposition and obesity, we initially compared NT-deficient mice[10,11] to wild-type littermates fed a standard diet (18% kcal from fat) for 6 months. Average body length, small intestine weight and length, villus height and crypt number did not differ between genotypes (Extended Data Fig. 1a–f). However, the epididymal and retroperitoneal fat pads of $Nt^{-/-}$ mice were consistently smaller as compared to wild type (Extended Data Fig. 2a–c). Indeed, when challenged with a high-fat diet (HFD; 60% kcal from fat), epididymal, retroperitoneal and pericardiac fat deposits were markedly smaller in $Nt^{-/-}$ mice (Extended Data Fig. 2d–f). Moreover, body weight was significantly less in both male and female $Nt^{-/-}$ mice fed a HFD (Fig. 1a, b) and male $Nt^{-/-}$ mice fed a low-fat diet (LFD; 10% kcal from fat) (Extended Data Fig. 2g) compared with wild type. No differences in body weight were noted between female $Nt^{+/+}$ and $Nt^{-/-}$ mice fed a LFD (Extended Data Fig. 2g). EchoMRI assessment of total fat (corrected for body weight) further confirmed reduced body fat composition in male $Nt^{-/-}$ mice (Extended Data Fig. 2h).

Obesity-associated insulin resistance was also attenuated in NT-deficient mice. On a HFD, NT-deficient mice demonstrated lower levels of fasting plasma glucose and insulin (Fig. 1c), greater insulin sensitivity and faster glucose clearance (Fig. 1d) compared with wild type. Consistently, insulin-stimulated phosphorylated (p)-Akt expression was decreased in the livers of $Nt^{+/+}$ mice fed a HFD, whereas $Nt^{-/-}$ mice fed the same diet demonstrated a similar induction of p-Akt after injection of insulin, as noted for both wild-type and NT-deficient mice fed a LFD (Extended Data Fig. 2i). In contrast, there was no difference in insulin secretion when comparing $Nt^{+/+}$ and $Nt^{-/-}$ mice maintained on a normal chow diet following either glucose administration by gavage or refeeding after a 16 h fast (Extended Data Fig. 2j, k). Hepatic steatosis (Fig. 1e), liver triglyceride (TG) and cholesterol accumulation (Fig. 1f) were significantly decreased in $Nt^{-/-}$ mice fed a HFD. Adipocytes in epididymal fat pads of $Nt^{-/-}$ mice showed a reduction in size (Fig. 1g, h) and decreased inflammatory infiltrates and numbers of macrophages (F4/80-positive cells) (Extended Data Fig. 2l, m). Together, these results indicate that NT deficiency protects against comorbidities (that is, increased insulin resistance and hepatic steatosis) associated with high dietary fat intake.

NT has been linked to hypothalamic leptin signalling[12] and is considered to be an anorectic peptide based on acute suppression of food intake in rats after intracerebral or intraperitoneal (i.p.) administration of NT[13,14]. When considering total food intake over 22 weeks, there were no differences between genotypes in either male or female mice fed a LFD or males fed a HFD; however, a slight 10% decrease in female $Nt^{-/-}$ mice fed a HFD reached significance (Extended Data Fig. 3a). Weekly food consumption was not statistically different in male or female mice fed a LFD or males fed a HFD (Extended Data

[1]Department of Surgery, University of Kentucky, Lexington, Kentucky 40536, USA. [2]Markey Cancer Center, University of Kentucky, Lexington, Kentucky 40536, USA. [3]Department of Toxicology and Cancer Biology, University of Kentucky, Lexington, Kentucky 40536, USA. [4]Department of Pharmacology and Nutritional Sciences, University of Kentucky, Lexington, Kentucky 40536, USA. [5]Department of Clinical Sciences, Lund University, Malmö, 221 00 Lund, Sweden. [6]Department of Internal Medicine, Skåne University Hospital, Malmö, 205 02 Malmö, Sweden. [7]Department of Molecular and Cellular Biochemistry, University of Kentucky, Lexington, Kentucky 40536, USA. [8]Center for Structural Biology, University of Kentucky, Lexington, Kentucky 40536, USA. [9]Center for Environmental and Systems Biochemistry, University of Kentucky, Lexington, Kentucky 40536, USA. [10]Division of Cardiovascular Medicine, Gill Heart Institute, University of Kentucky and Lexington Veterans Affairs Medical Center, Lexington, Kentucky 40536, USA. [11]Department of Biostatistics, University of Kentucky, Lexington, Kentucky 40536, USA. [12]Department of Microbiology and Physiological Systems, University of Massachusetts Medical School, Worcester, Massachusetts 01655, USA.
*These authors contributed equally to this work.
§These authors jointly supervised this work.

© 2016 Macmillan Publishers Limited. All rights reserved

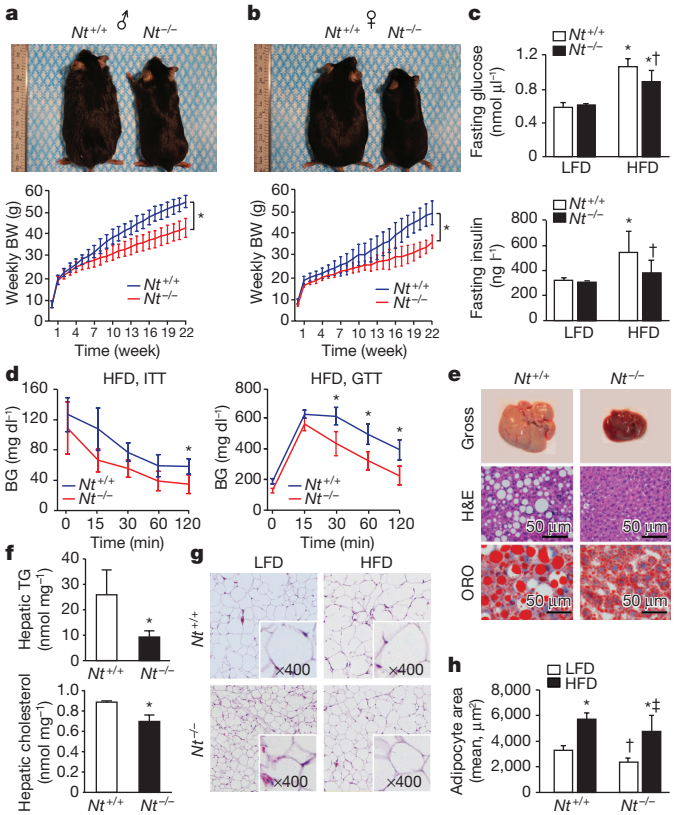


Figure 1 | Protective effects of NT deficiency on obesity and comorbid conditions. a, **b**, Representative male (**a**) and female (**b**) $Nt^{+/+}$ and $Nt^{-/-}$ mice fed a HFD for 22 weeks (top). Body weight (BW) was measured weekly (bottom) (male $Nt^{+/+}$ $n=18$ and $Nt^{-/-}$ $n=17$ mice; female $Nt^{+/+}$ $n=15$ and $Nt^{-/-}$ $n=12$ mice). Body weight slopes were compared: $^{*}P<0.05$ $Nt^{+/+}$ versus $Nt^{-/-}$ mice. **c**, Plasma glucose (top) and insulin (bottom) levels quantified in 24-h-fasted male mice maintained on a LFD or HFD for 24 weeks (LFD $Nt^{+/+}$ $n=10$, $Nt^{-/-}$ $n=10$; HFD $Nt^{+/+}$ $n=8$, $Nt^{-/-}$ $n=10$). $^{*}P<0.05$ versus LFD in $Nt^{+/+}$ and $Nt^{-/-}$ mice, respectively; $^{\dagger}P<0.05$ versus HFD in $Nt^{+/+}$ mice. **d**, Blood glucose (BG) during insulin tolerance test (ITT) (left; $n=5$) and glucose tolerance test (GTT) (right; $n=3$) in 6-h-fasted male mice fed a HFD for 24 weeks. $^{*}P<0.05$ versus $Nt^{-/-}$ mice. **e**, Gross, haematoxylin and eosin (H&E) and oil red O (ORO) imaging of livers of male mice fed a HFD for 24 weeks ($n=5$). Scale bar, 50 µm. **f**, Hepatic TG and cholesterol were analysed by liquid chromatography–mass spectrometry (LC–MS) in 24-h-fasted male mice fed a HFD for 24 weeks ($n=3$). $^{*}P<0.05$ versus $Nt^{+/+}$ mice. **g**, H&E staining of epididymal fat from male $Nt^{+/+}$ and $Nt^{-/-}$ mice fed a LFD or HFD for 24 weeks ($n=5$). **h**, Quantitative analysis of adipocyte area ($n=3$). $^{*}P<0.05$ versus LFD in $Nt^{+/+}$ and $Nt^{-/-}$ mice, respectively; $^{\dagger}P<0.05$ versus LFD in $Nt^{+/+}$ mice; $^{\ddagger}P<0.05$ versus HFD in $Nt^{+/+}$ mice. All data are mean ± standard deviation (s.d.). Linear mixed model for **a**, **b**; analysis of variance (ANOVA) with Holm's *P*-value adjustment for **c**, **h**; two-sided Student's *t*-test for **d**, **f** (see Methods).

Fig. 3b, c); only the week 9 comparison in females fed a HFD reached significance (Extended Data Fig. 3c). Energy expenditure, locomotor activity, energy intake and respiratory exchange ratios were not different between $Nt^{+/+}$ and $Nt^{-/-}$ mice fed either a HFD or LFD (Extended Data Fig. 3d–g). Although we cannot completely dismiss a potential effect of NT deficiency on feeding behaviour, this factor alone does not appear to play a major part in the lower weight gain observed in $Nt^{-/-}$ mice fed a HFD.

We next evaluated differences in intestinal lipid absorption as a possible mechanism for the decreased weight gain noted in NT-deficient mice. Compared to $Nt^{+/+}$ mice, faecal TG content was increased by ~25% in $Nt^{-/-}$ mice fed a HFD (Fig. 2a), but was not associated with a change in stool colour, consistency or output (Extended Data Fig. 4a), indicating that NT deficiency decreases lipid absorption without overt signs of fat malabsorption (that is, steatorrhoea). Consistently, fewer and smaller lipid droplets were noted in the mucosa of the proximal intestine of $Nt^{-/-}$ mice at 30 and 60 min after olive oil gavage compared with $Nt^{+/+}$ mice (Fig. 2b). To track and quantify intestinal absorption better, $^{13}C_{18}$-oleic acid ($^{13}C_{18}$-OA) was administered by gavage and measured in the proximal intestine and plasma by Fourier-transform-based mass spectrometry (FT-MS). $^{13}C_{18}$-OA was significantly decreased in the intestine and at 2 and 3 h in the plasma of $Nt^{-/-}$ mice (Fig. 2c). NT administration (3,600 nmol kg^{-1} body weight, i.p.)[14] restored TG accumulation in $Nt^{-/-}$ mice to levels similar to those in $Nt^{+/+}$ mice (Fig. 2d). Similarly, pre-treatment of C57BL/6 mice with SR 48692 (2.5 mg kg^{-1}, i.p.), a selective non-peptide NTR1 antagonist that acts peripherally and centrally when administered either i.p. or by gavage[15], decreased intestinal fatty acid absorption after olive oil gavage (Fig. 2e). To demonstrate further that the effect of SR 48692 was due to disruption of NT signalling, we repeated the experiment using $Nt^{-/-}$ mice and their wild-type littermates and found that, similar to our initial results, pre-treatment with SR 48692 inhibited intestinal fatty acid absorption in wild-type mice (as measured by TG accumulation); however, SR 48692 pre-treatment in $Nt^{-/-}$ mice did not further decrease lipid absorption after olive oil administration, which, as expected, was significantly decreased compared to wild-type mice (Extended Data Fig. 4b). These results indicate that SR 48692 attenuation of fat uptake reflects the disruption of normal NT signalling and is not due to an unanticipated off-target effect. NT treatment also increased fatty acid uptake in rat intestinal epithelial-1 (RIE-1) cells (Extended Data Fig. 4c, d) that express NTR1 and 3 (similar to human intestinal cells (FHs 74 Int)) and mouse intestinal mucosa (Extended Data Fig. 4e–g), and short interfering RNA (siRNA) knockdown of either *NTR1* or *NTR3*—the protein products of which have been demonstrated to heterodimerize on the cell surface and to broaden the response range for NT signalling[15]—reduced NT-mediated fatty acid absorption (Extended Data Fig. 4h, i). Consistent with a role for NT in HFD-induced weight gain, treatment with SR 48692 (2.5 mg per kg body weight, oral gavage, twice a day) for 13 weeks significantly attenuated body weight gain in wild-type mice fed a HFD (Fig. 2f) without altering food intake (Extended Data Fig. 4j). Collectively, these results indicate that NT promotes intestinal lipid uptake through NTR1, and possibly NTR3, promoting weight gain in mice fed a HFD. Interestingly, it was recently demonstrated[16] that NTR3-deficient mice, when placed on a HFD, exhibited a similar phenotype as NT-deficient mice, thus further emphasizing the potential importance of the NT/NTR axis in weight gain associated with an overabundance of fat.

AMPK, a key fuel-sensing enzyme and a critical regulator of metabolism, mediates the effects of a variety of hormones[17,18]. Phosphorylated-AMPK (p-AMPK) was increased in the proximal intestinal mucosa of $Nt^{-/-}$ mice fed a standard diet compared with wild type (Extended Data Fig. 5a). p-AMPK expression was decreased in the proximal intestine of $Nt^{+/+}$ mice after olive oil gavage but was restored to control levels with SR 48692 pre-treatment (Extended Data Fig. 5b). Treatment of FHs 74 Int and RIE-1 cells with oleate led to an increase in p-AMPK (Extended Data Fig. 5c), which was reduced by NT pre-treatment (Fig. 2g and Extended Data Fig. 5d). Pharmacological activation of AMPK with 5-aminoimidazole-4-carboxamide-1-β-D-ribofuranoside (AICAR) further increased oleate-stimulated p-AMPK (Fig. 2h) and concomitantly decreased fatty acid absorption (Fig. 2i); these effects were blocked by NT pre-treatment. NT-mediated suppression of p-AMPK was prevented by either NTR1 or NTR3 knockdown (Extended Data Fig. 5e, f). Knockdown of the upstream AMPK kinase Ca^{2+}-calmodulin-dependent protein kinase kinase (CaMKK2)[19,20], but not liver kinase B1 (LKB1)[21], decreased oleate-stimulated p-AMPK (Extended Data Fig. 5g). Moreover, overexpression of CaMKK2 attenuated NT-mediated suppression of p-AMPK (Extended Data Fig. 5h). Together, these findings suggest that NT, acting through

© 2016 Macmillan Publishers Limited. All rights reserved

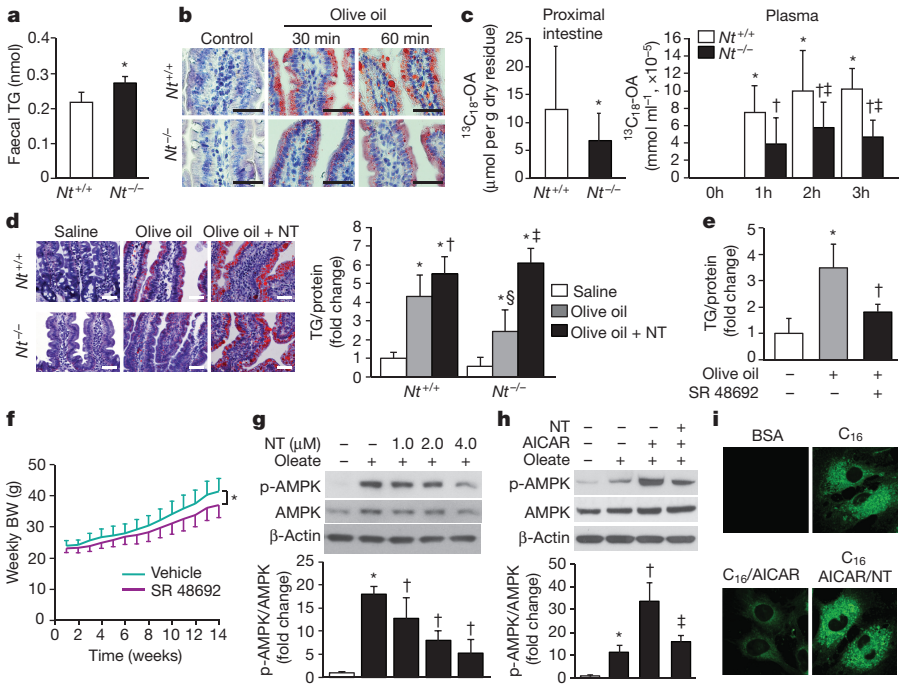


Figure 2 | NT deficiency reduces intestinal lipid absorption. a, Faecal TG was analysed in male mice fed a HFD for 24 weeks ($n = 3$). $*P = 0.05$ versus $Nt^{+/+}$ mice. **b**, ORO staining of proximal intestines from male mice fed normal chow with or without olive oil by oral gavage after overnight fasting ($n = 3$). Scale bar, 50 μm. **c**, Levels of $^{13}C_{18}$-OA in male mice fed normal chow and given $^{13}C_{18}$-OA mixed in olive oil by gavage after an overnight fast were analysed by nanospray FT-MS. Left ($n = 5$), $*P < 0.05$ versus $Nt^{+/+}$ mice. Right ($n = 4$), $*\dagger P < 0.05$ versus 0 h in $Nt^{+/+}$ and $Nt^{-/-}$ mice, respectively; $\ddagger P < 0.05$ versus 2 h and 3 h in $Nt^{+/+}$ mice. **d**, ORO staining of proximal intestines from male mice following gavage with saline, olive oil, or olive oil plus NT (3,600 nmol kg^{-1} body weight, i.p.) after an overnight fast (left; $n = 5$). Scale bar, 50 μm. TG (in mg) was quantified in proximal intestines and normalized to the amount of protein (in mg) as described in Methods (right; $n = 5$). Graph presents the fold change versus saline in $Nt^{+/+}$ mice. $*P < 0.05$ versus saline in $Nt^{+/+}$ and $Nt^{-/-}$ mice, respectively; $\dagger P < 0.05$ versus olive oil in $Nt^{+/+}$ mice; $\ddagger P < 0.05$ versus olive oil in $Nt^{-/-}$ mice; $\S P < 0.05$ versus olive oil in $Nt^{+/+}$. **e**, Proximal intestines from mice given saline, olive oil, or olive oil plus SR 48692 were collected and TG levels were quantified as described earlier ($n = 8$). Graph presents the fold change versus control mice. $*P < 0.05$ versus control mice; $\dagger P < 0.05$ versus mice with olive oil only. **f**, Weekly body weight was measured in male wild-type C57BL/6 mice fed a HFD and treated with SR 48692 (2.5 mg per kg body weight, diluted in distilled H_2O and administered by oral gavage twice a day) or vehicle (vehicle $n = 12$; SR 48692 $n = 13$). Body weight slopes were compared: $*P < 0.05$ vehicle versus SR treatment. **g**, FHs 74 Int cells were pre-treated with or without NT at different doses as indicated for 30 min followed by combined treatment with oleate (0.1 mM) or bovine serum albumin (BSA) for 1 h and western blotting of cell extracts (top); densitometric analysis of p-AMPK is from three separate experiments and normalized to total AMPK; graph demonstrates the fold change of p-AMPK versus control (BSA; bottom). $*P < 0.05$ versus BSA; $\dagger P < 0.05$ versus oleate alone. **h**, FHs 74 Int cells were treated with or without NT (2 μM) for 30 min followed by AICAR (1 mM) for 2 h and oleate (0.1 mM) or BSA for 1 h and analysed by western blot (top); p-AMPK levels were determined as in **g** from three separate experiments (bottom). $*P < 0.05$ versus control (BSA); $\dagger P < 0.05$ versus oleate alone; $\ddagger P < 0.05$ versus oleate plus AICAR. **i**, FHs 74 Int cells were treated with NT (2 μM) for 30 min followed by addition of AICAR (1 mM) for another 3 h; cells were then incubated with BSA or 4,4-difluoro-5,7-dimethyl-4-bora-3a,4a-diaza-*s*-indacene-3-hexadecanoic acid C_{16} (BODIPY FL C_{16}) for 15 min and images were taken by confocal microscopy (original magnification, ×180). Representative images are from three experiments. All data are mean ± s.d. Two-sided Student's *t*-test for **a**, **c** (left); ANOVA with Holm's *P*-value adjustment for **c** (right), **d** (right), **e**, **g** (bottom), **h** (bottom); linear mixed model for **f** (see Methods). (See Supplementary Fig. 1 for gel source data.)

NTR1 and/or NTR3, increases fatty acid absorption through suppression of CaMKK2-mediated AMPK phosphorylation.

Drosophila provide a powerful model system to understand molecular mechanisms regulating human metabolic disorders better. To establish further the role of NT on intestinal lipid absorption and AMPK regulation, human full-length *NT* complementary DNA (cDNA) was expressed in *Drosophila* midgut enteroendocrine cells using the enteroendocrine-cell-specific driver *Gr36C*-Gal4 (ref. 22) (see colocalization with enteroendocrine-cell-specific transcription factor Prospero[23,24]; Extended Data Fig. 6a, b); mature NT^{1-13} peptide was

Table 1 | Pro-NT levels and association with clinical outcomes

	N total/*N* cases*	Odds ratio (95% CI)				*P* for trend‡
		Pro-NT quartile 1†	Pro-NT quartile 2	Pro-NT quartile 3	Pro-NT quartile 4	
Prevalent obesity	4,626/604	1.0 (ref)	1.00 (0.78–1.29)	1.13 (0.88–1.45)	1.34 (1.05–1.70)	0.01
Prevalent abdominal obesity	4,625/1,769	1.0 (ref)	1.07 (0.90–1.27)	1.23 (1.04–1.46)	1.30 (1.09–1.54)	0.001
Prevalent insulin resistance	4,468/1,140	1.0 (ref)	1.30 (1.06–1.59)	1.43 (1.17–1.74)	1.70 (1.39–2.06)	<0.0001
New-onset obesity	2,594/333	1.0 (ref)	1.41 (0.95–2.10)	1.79 (1.21–2.65)	2.05 (1.38–3.06)	<0.0001

ref, reference group.

*Denotes total number of subjects in the analysis divided by the number of cases with listed obesity or insulin resistance phenotype. †Pro-NT quartiles 1–4 define the MDC-CC population quartiles (lowest to highest) with median (range) baseline fasting plasma pro-NT concentrations (pmol l^{-1}) of 60.1 (3.3–75.9), 89.3 (75.9–105), 123 (105–149) and 190 (149–1,155), respectively. Data are presented as odds ratios (95% CIs), and subjects belonging to the lowest quartile of pro-NT were defined as the reference group (odds ratio = 1). ‡Denotes the *P* value for linear trend over quartiles 1–4. Logistic regression models were used for comparisons between quartiles and linear trends over quartiles (see Methods).

© 2016 Macmillan Publishers Limited. All rights reserved

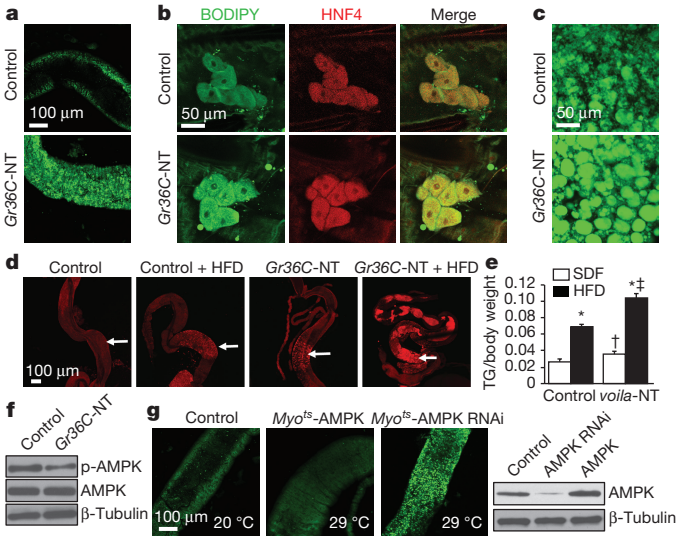


Figure 3 | NT suppresses AMPK activation and promotes lipid accumulation in *Drosophila*. **a**, Midguts from 7-day-old adult flies expressing either *Gr36C-w1118* (control, 100%, $n=7$; the percentage used here and in subsequent analyses indicates the percentage of organs exhibiting the phenotype) or *Gr36C*-NT (93%, $n=15$) were stained with BODIPY. Scale bar, 100 μm. **b**, Oenocytes on the basal surface of the lateral epidermis of 3rd instar larvae expressing *Gr36C-w1118* (100%, $n=15$) or *Gr36C*-NT (93%, $n=15$) were stained with BODIPY to monitor lipid accumulation (green) and the anti-HNF4 antibody to mark the oenocytes (red). Scale bar, 50 μm. **c**, Fat bodies attached to the salivary gland from the 3rd instar larvae expressing *Gr36C-w1118* (100%, $n=14$) or *Gr36C*-NT (100%, $n=8$) were stained with BODIPY. Scale bar, 50 μm. **d**, Flies expressing *Gr36C-w1118* or *Gr36C*-NT ($n=5$) were fed either SDF or HFD, and guts were stained with Nile Red to examine the accumulation of lipids (arrows). Scale bar, 100 μm. **e**, NT was expressed by *voila*-Gal4 and TG (in mg) was measured and normalized to body weight (in mg) in male adult flies fed either SDF or HFD ($n=3$). *voila-w1118* served as control. $*P<0.05$ versus SDF in control- and NT-expressing flies, respectively; $\dagger P<0.05$ versus SDF in control flies; $\ddagger P<0.05$ versus HFD in control flies. **f**, Western blotting was performed to monitor the levels of AMPK in adult midgut shown in **a**. **g**, *Myo1A*-Gal4 combined with *tub-Gal80^{ts}* (*Myo^{ts}*) does not express active Gal4 at 20 °C permissive temperature. Shown on the left is the midgut from a 7-day-old adult expressing *Myo^{ts}*-AMPK raised at 20 °C, stained with BODIPY, and used as a control (100%, $n=11$). Midguts expressing AMPK (middle; 94%, $n=16$) or AMPKRNAi^{25931} (right; 100%, $n=12$) from 7-day-old adults at 29 °C (active Gal4) were stained with BODIPY. Scale bar, 100 μm. The levels of AMPK were monitored by western blot. All data are mean ± s.d. ANOVA with Holm's *P*-value adjustment for **e** (see Methods). (See Supplementary Fig. 1 for gel source data.)

detected in larval gut (3.1 fg per gut) and *Drosophila* S2 cells transfected with *NT* cDNA (Extended Data Fig. 7). Compared to control, NT expression markedly increased lipid droplets in the midgut of 7-day-old adult (Fig. 3a) and larvae (Extended Data Fig. 6c) fed a standard diet, and also increased the accumulation of lipid droplets in oenocytes (Fig. 3b) and fat bodies (Fig. 3c) of *Gr36C*-NT 3rd instar larvae. Next, we used a diet-induced obesity model in *Drosophila*[25] to demonstrate increased lipid droplets in the midgut of control flies fed a HFD compared with flies fed standard food (SDF). NT expression dramatically increased lipid accumulation in the midgut (Fig. 3d) and total body TG levels (Fig. 3e) with either a SDF or HFD, suggesting that NT promotes efficient lipid absorption, and that this effect is further enhanced by increased fat concentration. Consistently, we also found that NT expression decreased gut p-AMPK levels in both adults and larvae (Fig. 3f and Extended Data Fig. 6d). These findings suggest that, similar to mice, the effects of NT on lipid absorption are mediated, in part, through AMPK regulation. Indeed, we found that AMPK overexpression decreased, whereas AMPK knockdown increased lipid droplets in the midgut of 7-day-old adults (Fig. 3g) and the midgut, fat body and oenocytes of larvae (Extended Data Fig. 6e–g).

To identify the endogenous *Drosophila* NTR, we carried out a targeted RNA interference (RNAi) screen in S2 cells; expression of NT or treatment with NT peptide consistently decreased p-AMPK (Extended Data Fig. 8a, b). Among the three potential *Drosophila* NTRs[26,27], RNAi of CG9918 (Pyrokinin 1 receptor (PK1-R)), but not CG8784 or CG8795, blocked the decrease in p-AMPK levels in NT-expressing cells (Extended Data Fig. 8a). The effect of NT does not appear to be due to interference with PK-1 signalling, since PK-1 (encoded by *Capability*) knockdown did not alter p-AMPK levels (Extended Data Fig. 8d). Knockdown of CG9918 expression in enterocytes using *Myo1A*-Gal4 to drive *UAS-CG9918 RNAi* (TRiP 27539) expression in flies also expressing NT driven by the enteroendocrine-cell-specific tachykinin (TK) promoter[28,29] markedly attenuated NT-induced lipid droplet accumulation (Extended Data Fig. 8c). These results indicate that NT increases lipid accumulation, at least in part, through CG9918, an NTR-like receptor that shares 32% identity and 50% similarity with mouse NTR1 (Extended Data Fig. 8e), suggesting an evolutionarily conserved function for NTR signalling in lipid uptake.

The mouse and *Drosophila* data prompted us to assess the possible role of NT in the development of obesity and its metabolic complications in humans. Fasting plasma concentrations of pro-NT were analysed from 4,632 middle-aged subjects of the population-based Malmö Diet and Cancer Study Cardiovascular Cohort (MDC-CC)[9] (Extended Data Table 1). The age- and sex-adjusted likelihood of being obese, abdominally obese and insulin resistant significantly increased across quartiles of pro-NT plasma levels ($P=0.01$, 0.001 and <0.0001, respectively; Table 1). Continuous values of pro-NT were also significantly related to continuous values of body mass index (BMI), waist circumference and homeostasis model assessment of insulin resistance (Extended Data Table 2). Among non-obese subjects, the risk of developing obesity during an average follow-up time of 16.5 ± 1.5 years increased gradually with pro-NT quartiles, independently of baseline BMI, age and gender ($P<0.0001$; Table 1). Importantly, non-obese subjects in the top quartile of baseline pro-NT levels had greater than double the risk of developing obesity compared to those in the lowest quartile (odds ratio = 2.05, 95% confidence interval (CI): 1.38–3.06). Whereas the cross-sectional relationship between pro-NT and obesity became non-significant after additional adjustment for insulin resistance, the prospective relationship between pro-NT and risk of new-onset obesity remained highly significant ($P=0.001$) after adjustment for insulin resistance (data not shown). Thus, pro-NT levels strongly predict new onset obesity in a graded manner, which is independent of baseline BMI and insulin resistance. These findings in adults warrant further prospective studies to evaluate whether pro-NT levels can be used to predict future obesity in children and adolescents.

Our findings demonstrate a critical role of NT in HFD-induced obesity that involves decreased AMPK activation and increased intestinal lipid absorption. Moreover, we identify increased pro-NT levels as a strong risk factor of human obesity. From an evolutionary perspective, metabolically 'thrifty' genes, such as NT, are highly beneficial to ensure the efficient absorption of all ingested fats, but with the abundance of fats in typical Western diets, NT can have a detrimental effect by contributing to increased fat storage, obesity and related metabolic disorders.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 24 February 2015; accepted 10 March 2016.
Published online 11 May 2016.

© 2016 Macmillan Publishers Limited. All rights reserved

1. Ogden, C. L., Yanovski, S. Z., Carroll, M. D. & Flegal, K. M. The epidemiology of obesity. *Gastroenterology* **132,** 2087–2102 (2007).
2. Kopelman, P. G. Obesity as a medical problem. *Nature* **404,** 635–643 (2000).
3. Guh, D. P. *et al.* The incidence of co-morbidities related to obesity and overweight: a systematic review and meta-analysis. *BMC Public Health* **9,** 88 (2009).
4. Polak, J. M. *et al.* Specific localisation of neurotensin to the N cell in human intestine by radioimmunoassay and immunocytochemistry. *Nature* **270,** 183–184 (1977).
5. Ferris, C. F., Hammer, R. A. & Leeman, S. E. Elevation of plasma neurotensin during lipid perfusion of rat small intestine. *Peptides* **2** (suppl. 2), 263–266 (1981).
6. Armstrong, M. J., Parker, M. C., Ferris, C. F. & Leeman, S. E. Neurotensin stimulates [3H]oleic acid translocation across rat small intestine. *Am. J. Physiol.* **251,** G823–G829 (1986).
7. Evers, B. M. Neurotensin and growth of normal and neoplastic tissues. *Peptides* **27,** 2424–2433 (2006).
8. Vincent, J. P., Mazella, J. & Kitabgi, P. Neurotensin and neurotensin receptors. *Trends Pharmacol. Sci.* **20,** 302–309 (1999).
9. Melander, O. *et al.* Plasma proneurotensin and incidence of diabetes, cardiovascular disease, breast cancer, and mortality. *J. Am. Med. Assoc.* **308,** 1469–1475 (2012).
10. Dobner, P. R., Fadel, J., Deitemeyer, N., Carraway, R. E. & Deutch, A. Y. Neurotensin-deficient mice show altered responses to antipsychotic drugs. *Proc. Natl Acad. Sci. USA* **98,** 8048–8053 (2001).
11. Piliponsky, A. M. *et al.* Neurotensin increases mortality and mast cells reduce neurotensin levels in a mouse model of sepsis. *Nature Med.* **14,** 392–398 (2008).
12. Sahu, A., Carraway, R. E. & Wang, Y. P. Evidence that neurotensin mediates the central effect of leptin on food intake in rat. *Brain Res.* **888,** 343–347 (2001).
13. Boules, M. *et al.* A novel neurotensin peptide analog given extracranially decreases food intake and weight in rodents. *Brain Res.* **865,** 35–44 (2000).
14. Cooke, J. H. *et al.* Peripheral and central administration of xenin and neurotensin suppress food intake in rodents. *Obesity (Silver Spring)* **17,** 1135–1143 (2009).
15. Martin, S., Navarro, V., Vincent, J. P. & Mazella, J. Neurotensin receptor-1 and -3 complex modulates the cellular signaling of neurotensin in the HT29 cell line. *Gastroenterology* **123,** 1135–1143 (2002).
16. Rabinowich, L. *et al.* Sortilin deficiency improves the metabolic phenotype and reduces hepatic steatosis of mice subjected to diet-induced obesity. *J. Hepatol.* **62,** 175–181 (2015).
17. Lim, C. T., Kola, B. & Korbonits, M. AMPK as a mediator of hormonal signalling. *J. Mol. Endocrinol.* **44,** 87–97 (2010).
18. Xue, B. & Kahn, B. B. AMPK integrates nutrient and hormonal signals to regulate food intake and energy balance through effects in the hypothalamus and peripheral tissues. *J. Physiol.* **574,** 73–83 (2006).
19. Hardie, D. G. AMPK: positive and negative regulation, and its role in whole-body energy homeostasis. *Curr. Opin. Cell Biol.* **33,** 1–7 (2015).
20. Racioppi, L. & Means, A. R. Calcium/calmodulin-dependent protein kinase kinase 2: roles in signaling and pathophysiology. *J. Biol. Chem.* **287,** 31658–31665 (2012).
21. Shackelford, D. B. & Shaw, R. J. The LKB1–AMPK pathway: metabolism and growth control in tumour suppression. *Nature Rev. Cancer* **9,** 563–575 (2009).
22. Park, J. H. & Kwon, J. Y. Heterogeneous expression of *Drosophila* gustatory receptors in enteroendocrine cells. *PLoS ONE* **6,** e29022 (2011).
23. Micchelli, C. A. & Perrimon, N. Evidence that stem cells reside in the adult *Drosophila* midgut epithelium. *Nature* **439,** 475–479 (2006).
24. Ohlstein, B. & Spradling, A. The adult *Drosophila* posterior midgut is maintained by pluripotent stem cells. *Nature* **439,** 470–474 (2006).
25. Birse, R. T. *et al.* High-fat-diet-induced obesity and heart dysfunction are regulated by the TOR pathway in *Drosophila*. *Cell Metab.* **12,** 533–544 (2010).
26. Hewes, R. S. & Taghert, P. H. Neuropeptides and neuropeptide receptors in the *Drosophila melanogaster* genome. *Genome Res.* **11,** 1126–1142 (2001).
27. Park, Y., Kim, Y. J. & Adams, M. E. Identification of G protein-coupled receptors for *Drosophila* PRXamide peptides, CCAP, corazonin, and AKH supports a theory of ligand-receptor coevolution. *Proc. Natl Acad. Sci. USA* **99,** 11423–11428 (2002).
28. Song, W., Veenstra, J. A. & Perrimon, N. Control of lipid metabolism by tachykinin in *Drosophila*. *Cell Reports* **9,** 40–47 (2014).
29. Amcheslavsky, A. *et al.* Enteroendocrine cells support intestinal stem-cell-mediated homeostasis in *Drosophila*. *Cell Reports* **9,** 32–39 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank D. A. Gilbreath, C. E. Anthony, H. N. Russell-Simmons and J. F. Rogers for manuscript preparation; D. Napier for tissue sectioning and staining; G. Epperly for intestinal crypt measurements and the use of the Aperio ScanScope; E. Y. Lee for consultation and assessment of histological sections and immunohistochemistry; R. Carraway and P. Forgez for their suggestions and helpful advice; J. Ambati and K. Grzech for their review of the manuscript; J. Young Kwon for the *Gr36C*-Gal4 line, M. Vidal for the *voila*-Gal4 line, N. Perrimon and T. Ip for the *TKg*-Gal4 line, and S. Hou for *MyoIA*-Gal4 and *Esg*-Gal4 lines; the Bloomington Stock Center, Vienna *Drosophila* RNAi Center (VDRC) and TRiP at Harvard Medical School for fly stocks; the Developmental Studies Hybridoma Bank (DSHB) for antibodies; B. Gebelein for the anti-HNF4 antibody; the Biospecimen and Tissue Procurement, Redox Metabolism, and Biostatistics and Bioinformatics Shared Resource Facilities of the University of Kentucky Markey Cancer Center (supported by National Cancer Institute grant P30 CA177558). This study was further supported by National Institutes of Health (NIH) grants R37 AG10885 and R01 DK048498 to B.M.E.; R01 GM079684 to J.J.; U24 DK097215, R01 ES022191, P01 CA163223 and P20 GM103527 to T.W.-M.F. and R.M.H.; R01 HL120507 to A.J.M.; RO1 NS077284 to H.Z.; R01 HL073085 and P20 GM103527 to L.A.C. O.M., P.M.N. and M.O.-M. are funded by the Swedish National Research Council; the Swedish Heart-Lung Foundation; Novo Nordisk Foundation; Swedish Diabetes Association; Region Skåne, ALF; and European Research Council grant StG-2011-282255. B.M.E. is also supported by the Markey Cancer Foundation. Y.Y.Z. and J.W.H. are supported by NIH postdoctoral training grants T32 CA165990 and T32 CA160003, respectively. The LC-MS/MS equipment was acquired using a National Center for Research Resources High-End Instrumentation grant (S10 RR029127 to H.Z.).

Author Contributions J.L., J.J., P.R.D., O.M. and B.M.E. designed the research; J.L., J.S., Y.Y.Z., Y.L., P.R. and K.J. performed experiments; J.T.K. assisted with experiments; M.E.S. and J.W.H. assisted with the animal work; W.S.K., F.B.Y. and L.A.C. performed indirect metabolism and adipocyte size analyses; J.C. and H.Z. designed and performed the proteomics studies using liquid chromatography tandem mass spectrometry (LC–MS/MS) for NT analysis of the *Drosophila* samples; A.J.M. performed LC–MS analysis; T.W.-M.F. and R.M.H. designed the ^{13}C-OA tracer study, reviewed the manuscript, and together with T.F. performed FT-MS data analyses; H.L.W. performed statistical analyses; P.R.D. provided NT-knockout mice and reviewed the manuscript; T.G. reviewed the manuscript and provided comments and suggestions; P.M.N., M.O.-M. and O.M. performed and analysed human studies; J.L., J.J., H.L.W. and B.M.E. reviewed and interpreted data; J.L., J.J., P.R.D. and B.M.E. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.M.E. (mark.evers@uky.edu).

© 2016 Macmillan Publishers Limited. All rights reserved

METHODS

Reagents. Phospho-AMPKα (Thr172) (2535), AMPKα (2532), LKB1 (3050), phospho-Akt (Ser-473) (4058) and pan-Akt (4691) antibodies were from Cell Signaling Technology (Danvers, MA). NTR1 antibody (sc-374492) was from Santa Cruz Biotechnology (Dallas, TX). CaMKK2 (ab168818), NTR3 (ab16640), and F4/80 (ab100790) antibodies were from Abcam (Cambridge, MA). Flag (F1804) and β-actin (A5316) antibodies were from Sigma-Aldrich (St Louis, MO). AICAR was from Cayman (Ann Arbor, MI). Oleate sodium, NT_{1-13}, glucose and human insulin were from Sigma. Deuterated oleic acid (CLM-460-PK) was obtained from Cambridge Isotope Laboratories (Tewksbury, MA). SR 48692 was from Tocris (Minneapolis, MN). Lipofectamine RNAiMAX and LTX Reagent with PLUS transfection reagents and Trizol were from Life Technologies (Grand Island, NY). pSG5-Flag-CaMKK2 rat FL was a gift from A. Means[30] (Addgene plasmid #32449, Cambridge, MA). Primers for RT–PCR were from Integrated DNA Technologies (Coralville, IA).

siRNA and sequences. ON-TARGETplus SMARTpool and ON-TARGETplus Non-targeting Control Pool siRNAs were purchased from GE Dharmacon (Lafayette, CO). The sequences are as below. Human *LKB1*: (1) UGACUGUGG UGCCGUACUU, (2) GCUCUUACGGCAAGGUGAA, (3) UGAAAGGGAUGCU UGAGUA, and (4) GAAGAAGGAAAUUCAACUA. Human *CAMKK2*: (1) GUGAAGACCAUGAUACGUA, (2) GGAUCGUGGUGCCGGAAAU, (3) GAUCAAAGGCAUCGAGUAC, and (4) ACAGUAAGAUCAAGAGUCA. Human *NTR1*: (1) GGACUCCGUUCCUCUAUGA, (2) GCAACACGGUGAC GGCGUU, (3) GAGCACAGCACAUUCAGCA, and (4) GAACACCGACAU CUACUCC. Human *NTR3* (also known as *SORT1*): (1) GAGACUAUGUUGUG ACCAA, (2) GAGCUAGGUCCAUGAAUAU, (3) GAAGGACUAUACCAUAUGG, and (4) GAAUUUGGCAUGGCUAUUG. Rat *Ntr1*: (1) GGGCACACACAA CGGUUUA, (2) CCGAAAUGGAAGCGACGUU, (3) CUACGUUCCUCUUCG AUUU, and (4) GCUACUAUUUCCUGCGUGA. Rat *Ntr3* (also known as *Sort1*): (1) ACAAAUGGGUACCGGAAAA, (2) GAACACAGCAACCGUCCUA, (3) AAGACAUCCUUGAGCGCAA, and (4) AAGCAGAAUUCCAAGUCGA.

Mice. All procedures were approved by the Institutional Animal Care and Use Committee of the University of Kentucky. $Nt^{-/-}$ mice and their wild-type littermates ($Nt^{+/+}$) were bred from $Nt^{+/-}$ mice and randomly grouped for all experiments. Mice were maintained with a 14 h light/10 h dark cycle and provided with food and water *ad libitum*. For diet-induced obesity studies, male and female $Nt^{+/+}$ and $Nt^{-/-}$ mice were placed on a 60% HFD or 10% LFD (catalogue no. D12492 and D12450B, respectively; Research Diets, New Brunswick, NJ) at weaning for 22–24 weeks. Body weight (and food intake were measured weekly. Food intake was measured for each cage (3–5 mice per cage) and divided by mouse number to obtain total grams consumed per mouse per week. All mice used were 4–6 months old unless otherwise indicated. For chronic SR 48692 treatment on HFD-fed mice, male C57BL/6 mice (2 months old) were obtained from Taconic. After 1 week acclimation, mice were placed in individual cages, started on HFD, and after 1 week were divided into two groups, one of which received SR 48692 (dissolved in sterile distilled H_2O by brief sonication, 2.5 mg per kg body weight[31–33] and the other vehicle twice a day (8 a.m. and 8 p.m.) by gavage. Body weight and food intake were measured weekly.

Human intestinal samples. Surgical samples of duodenum and colon were obtained from de-identified donors through the Markey Cancer Center Biospecimen and Tissue Procurement Shared Resource Facility. All samples were obtained after informed consent according to a protocol approved by the Institutional Review Board of the University of Kentucky Medical Center (UKMC). Tissues were processed within 1 h after resection; the mucosal layer was sharply dissected from the underlying seromuscular layer and collected in multiple 3–5 mm sections for western blotting.

Cell lines, transfection and treatment. FHs 74 Int human small intestinal epithelial cells[34] were purchased from ATCC (Manassas, VA) and maintained in Hybri-Care Medium ATCC 46-X supplemented with 30 ng ml^{-1} epidermal growth factor (Sigma-Aldrich) and 10% FBS. RIE-1 rat intestinal epithelial cells[35] were maintained in DMEM containing 2 mM L-glutamine, 4,500 mg l^{-1} glucose and 10% FBS. FHs 74 Int cells were tested for authentication via STR profiling in April 2015 by Genetica DNA Laboratories (LabCorp Speciality Testing Group; Burlington, NC) using the commercially available PowerPlex 16HS amplification kit (Promega Corporation) and GeneMapper ID v.3.2.1 software (Applied Biosystems). Authentication was confirmed by a 100% match in comparison to the reference STR profile from ATCC (FHs 74 Int; ATCC CCL-241). The cell lines are not listed in the International Cell Line Authentication Committee (ICLAC) database. In addition, both cell lines were tested for mycoplasma contamination via PCR (e-Myco Plus kit; iNtRON Biotechnology) and were found to be negative. Reverse transfection was performed using RNAiMAX (for siRNA) or LTX with PLUS (for plasmid) transfection reagents. Final siRNA concentrations were used at 20 (*NTR3*), 40 (*LKB1*, *CAMKK2*) or 100 (*NTR1*) nM. Cells were treated 72 h (siRNA) or 48 h (plasmid) after transfection. For combined treatment of oleate with NT, cells were pre-treated with BSA or NT (2 μM) or NT at various dosages in serum-free media as indicated in the figures for 30 min followed by oleate (0.1 mM) for 1 h or overnight. For treatment with oleate alone, cells were treated with BSA or oleate at different concentrations as indicated in the figures for 1 h in serum-free media.

Western blotting. Tissues and cells were lysed with lysis buffer (Cell Signaling Technology), and equal amounts of protein were resolved on 4–12% NuPAGE BisTris gels (Life Technologies), electrophoretically transferred to polyvinylidene difluoride (PVDF) membranes, and western blotting was performed as previously described[36,37]. p-AMPK expression was analysed by densitometry and normalized to total AMPK expression using NIH ImageJ software from three separate experiments. Data are presented as fold change.

Histology and immunohistochemistry. Tissues were fixed in 10% neutral-buffered formalin, embedded in paraffin, and sectioned (5 μm). H&E staining was performed using standard techniques. For H&E and oil red O (ORO) staining, liver tissues were collected from male $Nt^{+/+}$ and $Nt^{-/-}$ mice fed a HFD for 24 weeks after weaning. For ORO staining, liver tissues were fixed in 10% neutral-buffered formalin, equilibrated in 30% sucrose, embedded in OCT compound and snap-frozen in liquid N_2. Frozen sections were stained with ORO (Sigma-Aldrich) for lipid deposition using standard methods. Adipocyte size was measured as described previously[38]. Briefly, sections of epididymal adipose tissue from each mouse were photographed under ×100 magnification. In a square measuring 700 × 700 μm (*x* and *y* axis, respectively), adipocyte size and number were measured using NIS Elements BR.3.10 software. A criterion for inclusion of measurements was a circularity of adipocytes superior to 0.33 (shape of cells; from 0 (thin shape) to 1 (perfect circle)). Immunohistochemistry was performed and visualized by Dako EnVision Systems (Burlington, Ontario, Canada) following the product instruction.

RT–PCR. Total RNA was isolated from cells using RNeasy Kit (Qiagen, Valencia, CA) according to the manufacturer's instructions. RT–PCR analysis of *NTR1*, *2* and *3* was performed using cDNA synthesized from 1 μg of total RNA. β-Actin was used as the internal control. The primers included: human *NTR1*: 5′-TCATCGCCTTTGTGGTCTGCT-3′ and 5′-TGGTTGCTGGACACGCT GTCG-3′; human *NTR2*: 5′- GTCTCCTCAGCTTCATCGTAT-3′ and 5′-TCCCC AAAGCCTGAAGCTGTA-3′; human *NTR3*: 5′-AGAATGGTCGAGACTAT GTTG-3′ and 5′-AAGAGCTATTCCAAGAGGTCC-3′; rat *Ntr1*: 5′-GAGAAG CCCCCAAAATTCTC-3′ and 5′-CAAGGACCCAGTGCAGGTAT-3′; rat *Ntr2*: 5′-ACTCGCTCATCTTCGCATTT-3′ and 5′-TGGGACCACACGAAGTTGTA-3′; rat *Ntr3*: 5′-TTTCAAGCTGTGCTTTGTGG-3′ and 5′-AGTTCTCTGA ACGGGAGCAA-3′. β-Actin: 5′-TCACCAACTGGGACGACATG-3′ and 5′-ACCGGAGTCCATCACGATG-3′. The PCR products were analysed on a 2% agarose gel.

Glucose and insulin tolerance tests. Insulin (ITT) and glucose (GTT) tolerance tests were performed on 6-h-fasted male $Nt^{+/+}$ and $Nt^{-/-}$ mice fed HFD for 24 weeks after weaning. Glucose values were measured using One Touch Ultra from LifeScan (Wayne, PA) by tail snip. Glucose (1.5 g per kg body weight) and human insulin (0.75 U per kg body weight) were injected intraperitoneally (i.p.) after baseline glucose levels were established in each mouse, and blood glucose levels were measured 15, 30, 60 and 120 min after injection.

Metabolic studies. Whole-body composition parameters were measured in male $Nt^{+/+}$ and $Nt^{-/-}$ mice fed either a LFD or HFD for 24 weeks after weaning by EchoMRI-100 Whole Body Composition (Echo Medical System, Houston, TX) using magnetic resonance relaxometry to precisely measure total body fat, lean mass, body fluids and total body water in conscious mice. A TSE LabMaster indirect calorimetry system (TSE-Systems, Chesterfield, MO) was used to simultaneously quantify energy expenditure, energy intake, locomotor activity and respiratory exchange ratio (RER). Mice were acclimated to the chambers for 7 days, to permit recovery from the weight loss initially experienced by obese mice. Recordings were performed for 5 days, yielding three full 24-h periods of data. Feeding and activity data were collected continuously; O_2 and CO_2 levels for energy expenditure and RER calculations were collected at 30-min intervals. Resting energy expenditure values were calculated from data collected between 9 a.m. and 6 p.m., filtered to remove points at which activity was greater than 150 counts for that interval.

For hepatic TG and cholesterol measurements, liver tissues were collected from male $Nt^{+/+}$ and $Nt^{-/-}$ mice fed a HFD for 24 weeks after weaning; TG and cholesterol were extracted as described previously[39] and analysed by LC–MS coupled with electrospray ionization tandem using stable isotope dilution[40] performed on AB Sciex 4000 Q-Trap instruments. Fasting plasma glucose was analysed using a Glucose Colorimetric Assay Kit II (BioVision) and insulin using a Mouse Insulin ELISA kit (Mercodia, Winston Salem, NC). To monitor insulin secretion in response to glucose stimulation, plasma insulin and glucose levels

© 2016 Macmillan Publishers Limited. All rights reserved

were measured in mice (fed standard chow after weaning) that had been fasted overnight and subsequently given glucose (2 g per kg body weight) by gavage (that is, oral GTT). Blood was collected from tail snip before (0 min) or 15, 30, 60 and 120 min after gavage. Insulin and glucose levels were measured using Glucose Colorimetric Assay Kit II and Mouse Insulin ELISA kit as described earlier. Plasma insulin levels were also determined in mice that were either fasted for 16 h or fasted for 16 h with refeeding for 4 h; blood was collected from the IVC. For faecal TG assay, mice were housed in individual cages for 4 days, day-4 faecal samples were dried and ground, lipid was extracted from 50 mg faeces, and TG was measured by LC–MS.

Body length measurements. Anal–nasal length was measured following isoflurane anaesthesia of male and female $Nt^{+/+}$ and $Nt^{-/-}$ mice (7 months old).

Faecal weight measurements. Male $Nt^{+/+}$ and $Nt^{-/-}$ mice fed either normal chow (NC) or a HFD for 24 weeks after weaning were separated into individual cages and faecal pellets were collected and weighed daily for 4 days. Faecal weights from day 4 were averaged for wild-type and NT-deficient mice.

Small intestinal characterization. The entire small intestine (SI), from the gastric pylorus to the ileocecal valve was dissected from anaesthetized male $Nt^{+/+}$ and $Nt^{-/-}$ mice (7 months old). The SI length was measured and then opened longitudinally, washed in cold saline to clear the luminal contents, dried briefly on a paper towel and weighed. The SI was divided 5 cm distal from the pyloric junction and the jejunoileum was divided into equal proximal and distal fragments. Proximal fragments were fixed in 10% neutral-buffered formalin for 24 h, and 'Swiss rolls'[41] were sectioned (5 μm) for H&E staining. H&E-stained sections were imaged on an Aperio ScanScope XT slide scanner at ×20. Villus height and crypt counts were determined using Aperio ImageScope v.11.2.0.780 software. Crypts in a 1-mm field were counted; 10 fields were analysed per section. Villus height was measured from 10 well-oriented villi on each slide.

Lipid absorption studies. For olive oil administration and ORO analysis, male $Nt^{+/+}$ and $Nt^{-/-}$ mice on NC were fasted overnight, fed olive oil (17 μl per g body weight) by gavage, and killed either before (control), or 30 min and 60 min after gavage. The intestine was resected from the ligament of Treitz to the ileocecal junction, divided into proximal, middle, and distal segments of equal length, washed with cold saline, and the proximal segment was processed for frozen sectioning and ORO staining using standard protocols.

For $^{13}C_{18}$-OA experiments, male $Nt^{+/+}$ and $Nt^{-/-}$ mice on NC were fasted overnight, fed $^{13}C_{18}$-OA (480 μg per g body weight) mixed in olive oil (10 μl per g body weight) by gavage, killed either 0 or 30 min later, and proximal intestinal segments were collected as described earlier. To detect plasma level of $^{13}C_{18}$-OA, mice were given $^{13}C_{18}$-OA as described earlier and blood was collected from tail snip either before (0 h), or 1, 2 and 3 h after gavage. Lipid was extracted from proximal intestines and plasma[42] and $^{13}C_{18}$-OA levels were determined by direct infusion nanospray FT-MS (Orbitrap Fusion Tribrid Mass Spectrometer, Thermo Scientific, Waltham, MA) using a modification of a previous method[43]. Briefly, aliquots of the lipid extracts were diluted 20-fold in isopropanol:chloroform:methanol (4:2:1), introduced into the TriVersa NanoMate (Advion, Ithaca, NY) and analysed on the Orbitrap Fusion in negative-ion mode to estimate the appropriate amount of d_{34}-OA to spike. Lipid extracts were then spiked with a known concentration of d_{34}-OA and re-analysed. The nanospray conditions on the Nanomate were as follows: 15 μl of sample injection, 16 min of delivery time, 0.4 psi of gas pressure, and 1.5 kV of negative applied voltage. The Orbitrap Fusion MS analysis conditions were as follows: mass resolution setting of 450,000 with lock mass using internal calibrant, scan range of m/z 150–1,600, percentage S-Lens RF Level of 60, 1.0×10^5 for AGC Target, 100 ms maximum injection time, and 10 averaged microscans. The ion transfer tube temperature was 275 °C, and the experimental mass accuracy was ± ~1 p.p.m. The unlabelled, ^{13}C-labelled, and d_{34}-OA were assigned based on their respective accurate mass of 281.24860, 299.30899, and 314.45574; unlabelled and ^{13}C-labelled OA were then quantified against the internal d_{34}-OA standard using their respective peak intensity.

For NT 'rescue' experiments, male $Nt^{+/+}$ and $Nt^{-/-}$ mice on NC were fasted overnight, fed saline or olive oil (17 μl per g body weight) by gavage, injected i.p. with saline or NT (3,600 nmol per kg body weight), and, after being killed 60 min later, proximal intestinal fragments were dissected as described earlier and further divided into proximal and distal halves that were either processed for ORO staining or used for TG quantification, respectively, as described previously with modifications[44]. Briefly, intestinal tissues were homogenized in 200 μl of 0.5% Tween-20 in PBS, heated for 5 min at 70 °C, homogenates were microcentrifuged (3 min), and TG measurements were performed on the supernatants using a Triglyceride Determination Kit (Sigma-Aldrich)[44] according to the manufacturer's instructions.

For *in vitro* fatty acid absorption or lipid accumulation studies, FHs 74 Int and RIE-1 cells (in serum-free medium) were incubated with either bovine serum albumin (BSA; fatty-acid-free) (Sigma-Aldrich) or BSA-conjugated BODIPY FL C_{16} (4,4-difluoro-5,7-dimethyl-4-bora-3a,4a-diaza-*s*-indacene-3-hexadecanoic acid) (Life Technologies) (5 μM) for 15 min and images were obtained using an FV1000 Olympus confocal microscope (Olympus, Tokyo, Japan) using a ×60, 1.35 numerical aperture (NA) oil objective, zoom 3 and analysed with Olympus FV10-ASW v.3.1b software. Florescence intensity was determined on a total of 30 cells from 3 fields (10 cells per field). Lipid accumulation, RIE-1 cells were pre-treated with or without NT for 30 min followed by combination treatment of BSA or oleate (0.1 mM) with or without NT overnight. Cells were homogenized and TG measured using Triglyceride Determination Kit as described earlier. Experiments were repeated at least three times.

For acute SR 48692 treatment, one experiment was performed on male wild-type C57BL/6 mice obtained from Taconic (Hudson, NY). Mice were acclimated for 1 week, randomly divided into control, olive oil, and olive oil plus SR 48692 groups and fasted overnight. SR 48692 was dissolved in dimethylsulfoxide (DMSO; 50 mg ml^{-1}) and diluted 1:100 in saline just before use. Mice were injected i.p. with SR 48692 (2.5 mg per kg body weight)[31,33] or vehicle 30 min before oral gavage of olive oil (10 μl per g body weight), killed after an additional 30 min, and the proximal intestine (see earlier) was used to determine TG content and for preparation of lysates for western blotting. This experiment was repeated in male $Nt^{+/+}$ and $Nt^{-/-}$ mice (7 months old), and proximal intestinal TG was measured as described earlier.

***Drosophila* studies.** Full-length human *NT* cDNA was inserted into an attB-UAST backbone[45] and transgenic lines were generated using an attP2 locus and PhiC31 integration system. *UAS-AMPK* and *UAS-AMPK-RNAi* (v106200) lines have been described[46,47]. *UAS-AMPK-RNAi* line (TRiP #25931, Bloomington Stock Center) produced similar phenotypes to *UAS-AMPK-RNAi* (v106200) and was used for most experiments. The *UAS-CG9918-RNAi* line (TRiP #27539, Bloomington Stock Center) has been described[48]. *Gr36C*-Gal4, *Myo1A*^{ts}-Gal4, *voila*-Gal4, and *S106*-Gal4 have been described[49–51] (FlyBase). *Myo1A*-Gal4 is specifically expressed in enterocytes[52]. To constitutively express NT in gut enteroendocrine cells, a gut enteroendocrine-cell-specific TK promoter (2.0 kb)[28,29] was cloned into attB-UAST lacking Gal4-binding sites followed by insertion of full-length *NT* cDNA and the resulting plasmid (TK-NT) was used to generate a transgenic line by insertion at the VK5 attP locus. To express NT in enteroendocrine cells and knock down CG9918 by RNAi in enterocytes, the following genotype was used: *yw*; *Myo1A-Gal4/CG9918RNAi*^{27539}; *TK-NT*/+. *tub-Gal80*^{ts} was used in combination with the *Myo1A*-Gal4 (resulting in *Myo*^{ts}) to suppress *Myo1A-Gal4*-driven expression[52], until shifting larvae to 29 °C (non-permissive temperature) 96 h after egg laying (AEL) or adult emerging. AMPK or AMPK RNAi were expressed in the fat body using the RU486-inducible GeneSwitch *S106*-Gal4 line. *S106*-AMPK or *S106*-AMPK-RNAi larvae (3rd instar) were grown on RU486- (200 μM, Mifepristone, Sigma-Aldrich) or vehicle-containing food and lipid droplet accumulation in fat body and oenocytes was examined by BODIPY (Invitrogen) and immunohistochemical (anti-HNF4 oenocyte marker[53]) staining. Midguts (3rd instar larvae, adult) were dissected, fixed in 4% formaldehyde in PBS for 20 min, permeabilized in 0.1% Triton X-100 in PBS, and processed for either immunohistochemistry using mouse anti-Pros (DSHB, MR1A) and rabbit anti-NT (Abcam, ab43833) primary antibodies or lipid staining using BODIPY, Nile Red, or ORO (larvae).

AMPK and p-AMPK were detected by western blot analysis using protein extracts from gastrointestinal tract (15 per sample) using anti-p-AMPK (Cell Signaling, #2535), anti-AMPK (Abcam, ab80039) and anti-β-tubulin control (DSHB, E7) primary antibodies and ECL as described.

Adult flies were fed a HFD or standard food (SDF, cornmeal yeast) as described[25]. Briefly, flies (5 days after emerging) were collected and maintained on SDF for an additional 5 days, split into two groups, and fed either SDF or SDF containing 20% (w/v) coconut oil (HFD) for a further 5 days, followed by dissection and lipid staining using BODIPY or Nile Red. *voila*-NT and *voila*-Gal4-w1118 control flies were fed a HFD or SDF as described earlier for TG measurement (25 flies per sample) as described[44]. Fluorescence signals were acquired on an Olympus confocal microscope and images processed with Olympus FV10-ASW v.3.1b.

Drosophila S2 cells[54] were transfected with *ub*-Gal4, UAST-NT, and individual dsRNAs[55] to inhibit the expression of NTR-like receptors CG9918 (nucleotides 11–570), CG8784 (nucleotides 72–478), CG8795 (nucleotides 203–343), Capa (nucleotides 1–455) or GFP control (nucleotides 6–606) using Effectene (Qiagen). Cells (6×10^6) were dissolved in lysis buffer (450 μl) 48 h post-transfection, microcentrifuged (12,000 r.p.m., 10 min), and supernatants were analysed by western blotting. For quantitative RT–PCR, cDNA was synthesized (SuperScript III First Strand Synthesis, Invitrogen) from 1.0 μg Trizol-extracted RNA and used for quantitative RT–PCR (95 °C, 30 s followed by 40 cycles, 95 °C, 5 s; 55 °C, 30 s; 72 °C, 15 s) using SYBR Green (Thermo-Fisher) and the ΔC_t method (StepOnePlus Real-Time PCR System, Applied Biosystems). Primer pairs were as follows: CG9918, 5′-GAGTTTCAACGGCGGAGGAA-3′, 5′-AGCAGAGGAAGAAGCACACC-3′; CG8784, 5′-GGCGTGCTGGGTAATCT TAT-3′, 5′-CAAAGGTTGTACAGCTCCTG-3′; CG8795, 5′-GCTACGCCCTC

© 2016 Macmillan Publishers Limited. All rights reserved

ATATTTATC-3′, 5′-GAGGTTATAGAGGTCCTGCG-3′; Capa, 5′-ATGAAATCT ATGTTGGTC-3′, 5′-CCAACGCGCGGGAAGGC-3′; actin, 5′-GCGTCGGTCAA TTCAATCTT-3′, 5′-AAGCTGCAACCTCTTCGTCA-3′. All S2 cell experiments were independently repeated at least three times.

NT enzyme immunoassay. Fifty microlitres of media from S2 cells transfected with *ub*-Gal4 and UAST-NT were used to measure NT levels using NT EIA Kit from Phoenix Pharmaceuticals (Burlingame, CA) as described previously[36,37].

LC–MS/MS analysis of NT. NT was analysed by liquid chromatography tandem mass spectrometry (LC–MS/MS) using an LTQ-Orbitrap mass spectrometer (Thermo Fisher Scientific, Waltham, MA) coupled with an Eksigent Nanoflex cHiPLC system (Eksigent, Dublin, CA) through a nano-electrospray ionization source[56]. NT in the NT EIA kit (Phoenix Pharmaceutical) was used as standard. S2 cells were transfected with *ub*-Gal4 and UAST-NT or the UAST vector control, and conditioned medium was harvested after 48 h for analysis. GI tracts (350 per genotype) were dissected from wild-type (*w1118*) or *Gr36C*-NT 3rd instar larvae, methanol extracted[57], extracts were dried, and dissolved in 20 µl 0.1% (v/v) formic acid in water for analysis. Samples were separated by reversed-phase cHiPLC (ChromXP C18 column, 75 µm I.D. × 15 cm length, Eksigent catalogue no. 804-00001; mobile phase A and B were 0.1% (v/v) formic acid in either water or acetonitrile, respectively; flow rate, 300 nl min^{-1}). LC–MS/MS data were acquired in an automated data-dependent acquisition mode consisting of an Orbitrap MS scan (300−1,800 *m*/*z*, 60,000 resolutions) followed by MS/MS for fragmentation of the seven most abundant ions with the collision-induced dissociation method.

MS/MS fragments corresponding to NT^{1-13} were identified by comparison with the NT standard, and quantified using the intensity of the NT^{3+} peak and NT standards (0.05, 0.1, 0.2, 0.5 fmol)[57].

Human population studies. The Malmö Diet and Cancer (MDC) study is a population-based, prospective epidemiological cohort of 28,449 men (born 1923–1945) and women (born 1923–1950) from the city of Malmö in southern Sweden who underwent baseline examinations between 1991 and 1996[58]. From this cohort, 6,103 persons were randomly selected to participate in the MDC Cardiovascular Cohort (MDC-CC), which was designed to investigate the epidemiology of carotid artery disease, between 1991 and 1994[59]. Fasted plasma samples at the baseline examination were available for analysis of pro-NT and successfully measured in a total of 4,632 participants in the MDC-CC. Of those, complete data were available for BMI in 4,626, for waist circumference on 4,625, and for estimated degree of insulin resistance using the homeostasis model assessment of insulin resistance (HOMA-IR) (fasting blood glucose concentration × fasting plasma insulin concentration/22.5)[60] in 4,468 participants. BMI was defined as body weight in kilograms divided by the square of height in meters and obesity as a BMI $\geq$ 30 kg m^{-2}. Abdominal obesity was defined as a waist circumference of $\geq$94 cm in males and $\geq$80 cm in females, according to the International Diabetes Federation definition[61]. Insulin resistance was regarded present in subjects belonging to the top 25% of HOMA-IR values in the MDC-CC. 'New-onset obesity' is defined as obesity development among non-obese MDC-CC participants who were re-examined and diagnosed with obesity after an average follow-up time of 16.5 ± 1.5 years. Pro-NT was measured in stored fasting plasma specimens (all samples were assayed in 2010) that were immediately frozen to −80 °C at the MDC-CC baseline examination using a chemiluminometric sandwich immunoassay to detect a pro-NT precursor fragment (pro-NT 1–117) as described previously[62]. Analyses of blood glucose and plasma insulin were carried out at the time of baseline examination at the Department of Clinical Chemistry, Malmö University Hospital, which is attached to a national standardization and quality control system[63]. Of the 4,626 subjects with baseline data on BMI and pro-NT, 2,900 subjects were re-examined with a new measurement of BMI after a mean follow-up of 16.5 ± 1.5 years. In analyses of incident obesity, we excluded 306 subjects who were obese already at the baseline examination, leaving a total of 2,594 non-obese subjects for analysis of pro-NT in relation to incident obesity. All participants gave written informed consent, and the study was approved by the Ethical Committee at Lund University, Lund, Sweden.

Statistical analysis. Descriptive statistics including means and standard deviations were calculated and represented using bar or line graphs. Linear mixed models were implemented to compare body weight and food intake levels over time as well as comparisons at each time point for other endpoints from *in vivo* experiments involving repeat measurements. Specifically, body weight growth curves were compared using linear mixed models with fixed effects for linear and quadratic terms for time and their interaction with genotype and random effects for the intercept and time factors. Contrast statements for the interaction terms in the model were used to assess differences in growth curves between genotypes. Likewise, linear mixed models were employed for body weight growth curve comparisons between vehicle versus SR 48692 with the same fixed and random effects terms as earlier along with baseline weight as a covariate. Total food consumption for 22 weeks was calculated and compared between genotypes and between vehicle versus SR 48692 treatment using two-sample *t*-tests. Furthermore, weekly food intake levels were analysed using linear mixed models. However, since there is no clear trend in weekly food intake over time, an overall comparison of trend was not employed. Instead, a linear mixed model with fixed effects for genotype, time and their interaction and a random effect for the intercept term was employed in order to perform individual comparisons at each time point. A *P*-value adjustment due to multiple testing at each time point was performed using the Holm's step-down procedure. We used the Akaike information criterion (AIC) to evaluate the goodness-of-fit of the above linear mixed models. Two-sided, two-sample Student's *t*-test was employed for cell culture and *in vivo* studies involving two independent groups. One-way or two-way analysis of variance models with test for interaction between two factors was employed for multiple group comparisons of diet (LFD, HFD), mouse genotype, feeding and fasting groups, different time points of measurement, varying doses of NT treatment, SR48692 and AICAR treatments with Holm's *P*-value adjustment for multiple pairwise comparisons. ANCOVA was employed to compare genotype and diet with adjustment for the confounding effect of mouse body weight for fat composition and energy metabolism endpoints (energy expenditure, locomotor activity, energy intake, respiratory ratio). Model building was performed to assess equality of slopes using two-way interactions between each of genotype and diet with the body weight covariate as well as genotype and diet interactions.

For the animal studies, sufficient sample sizes were used to provide at least 80% power to detect large effect sizes (1.0 to 3.0 mean differences in s.d. units) based on two-sided, two-sample *t*-tests with 5% significance level. Some experiments including body weight and food intake were measured repeatedly, thus affording larger statistical power. No replicate samples from *in vitro* studies were excluded in the analysis. All data from animal studies with measurement of study endpoints were included in the analysis. Experiments with slight differences in animal numbers per group were due to the number of animals that were successfully bred and not due to exclusion of certain animals and their data points. Mice within a cage were randomized to all groups in an experiment to ensure balance in treatment group assignments across all cages. The animals were randomly selected for group assignment without preference to size or other confounding factors. A different individual performed measurements on study endpoints to ensure blinding from group assignment. Furthermore, only animal IDs without information on group assignment were available to staff performing the endpoint evaluation. Parametric tests were used after evaluating distribution of data (for example, percentiles, mean and median levels), test for normality (for example, Kolmogorov–Smirnov test, if sufficient sample sizes) and test for homogeneity of variance assumptions across groups. Otherwise, data log transformation or nonparametric tests were used. Appropriateness of other statistical models including linear mixed models was evaluated for goodness-of-fit using the AIC and equality of slopes between groups was evaluated for the ANCOVA models.

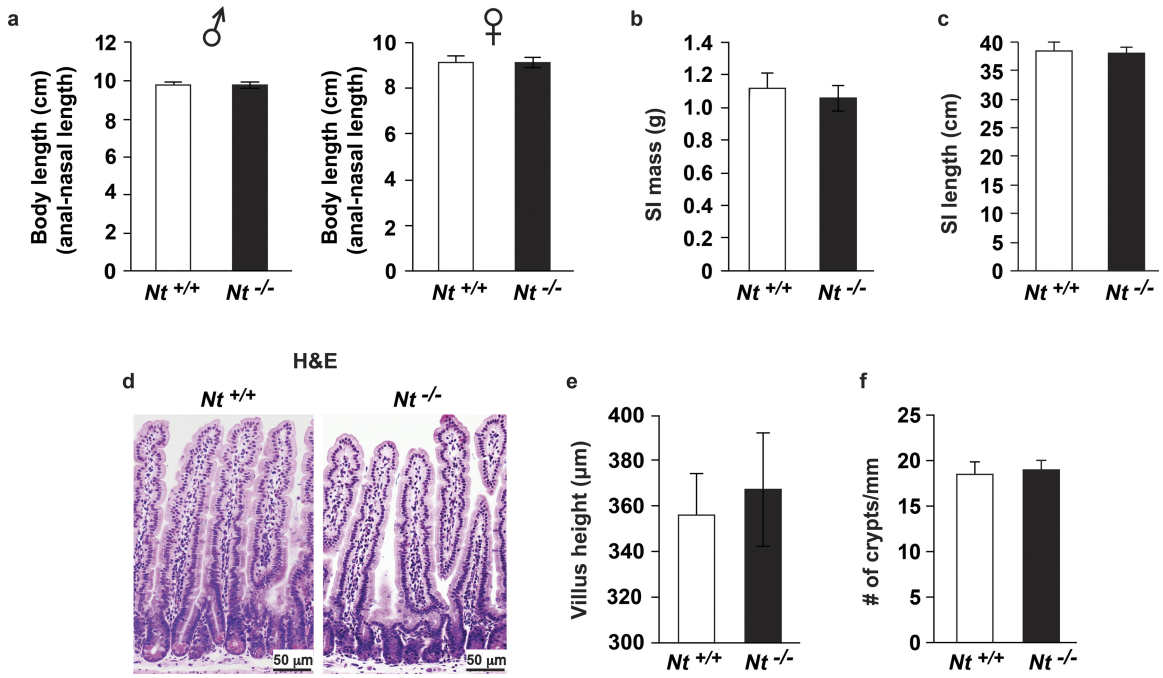
All subjects at the MDC-CC baseline examination were divided into ascending quartiles according to their value of fasting pro-NT. BMI, waist circumference and HOMA-IR were analysed as dichotomous outcome variables as defined earlier. In cross-sectional analyses, we related baseline quartile of pro-NT to the dichotomous outcome of obesity, abdominal obesity, and insulin resistance using age- and sex-adjusted logistic regression models. In the analyses of incident obesity, we related baseline quartile of pro-NT to the dichotomous outcome of incident obesity using logistic regression adjusted for baseline age, sex and BMI. Data are presented as odds ratios (95% confidence intervals), and subjects belonging to the lowest quartile of pro-NT were defined as the referent group (odds ratio = 1). '*P* for trend' denotes the *P* value for linear trend over quartiles 1–4. For code availability, statistical analysis codes in SAS version 9.4 were used for analysis of *in vivo* and *in vitro* studies and SPSS version 22.0 for analysis of human data and can be accessed by contacting B.M.E.

30. Anderson, K. A. *et al.* Hypothalamic CaMKK2 contributes to the regulation of energy balance. *Cell Metab.* **7,** 377–388 (2008).
31. Fadel, J., Dobner, P. R. & Deutch, A. Y. Amphetamine-elicited striatal Fos expression is attenuated in neurotensin null mutant mice. *Neurosci. Lett.* **402,** 97–101 (2006).
32. Steinberg, R. *et al.* SR 48692, a non-peptide neurotensin receptor antagonist differentially affects neurotensin-induced behaviour and changes in dopaminergic transmission. *Neuroscience* **59,** 921–929 (1994).
33. Fadel, J., Dobner, P. R. & Deutch, A. Y. The neurotensin antagonist SR 48692 attenuates haloperidol-induced striatal Fos expression in the rat. *Neurosci. Lett.* **303,** 17–20 (2001).
34. Owens, R. B., Smith, H. S., Nelson-Rees, W. A. & Springer, E. L. Epithelial cell cultures from normal and cancerous human tissues. *J. Natl. Cancer Inst.* **56,** 843–849 (1976).
35. Song, J., Li, J., Lulla, A., Evers, B. M. & Chung, D. H. Protein kinase D protects against oxidative stress-induced intestinal epithelial cell injury via Rho/ROK/PKC-δ pathway activation. *Am. J. Physiol. Cell Physiol.* **290,** C1469–C1476 (2006).

© 2016 Macmillan Publishers Limited. All rights reserved

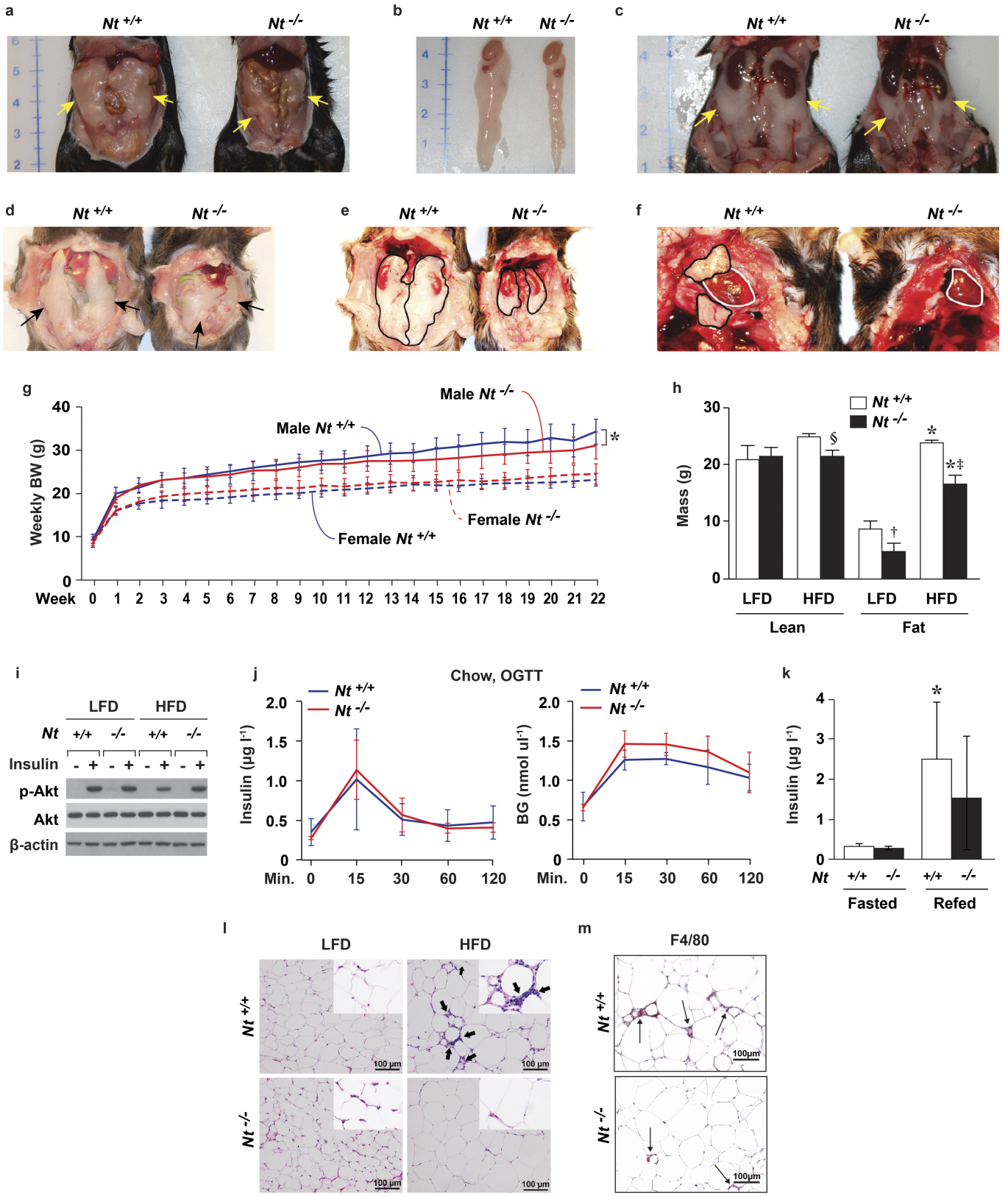
36. Li, J. *et al.* mTORC1 inhibition increases neurotensin secretion and gene expression through activation of the MEK/ERK/c-Jun pathway in the human endocrine cell line BON. *Am. J. Physiol. Cell Physiol.* **301,** C213–C226 (2011).
37. Li, J. *et al.* Cyclic adenosine 5′-monophosphate-stimulated neurotensin secretion is mediated through Rap1 downstream of both Epac and protein kinase A signaling pathways. *Mol. Endocrinol.* **21,** 159–171 (2007).
38. Yiannikouris, F. *et al.* Adipocyte-specific deficiency of angiotensinogen decreases plasma angiotensinogen concentration and systolic blood pressure in mice. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **302,** R244–R251 (2012).
39. Salous, A. K. *et al.* Mechanism of rapid elimination of lysophosphatidic acid and related lipids from the circulation of mice. *J. Lipid Res.* **54,** 2775–2784 (2013).
40. Onono, F. *et al.* Efficient use of exogenous isoprenols for protein isoprenylation by MDA-MB-231 cells is regulated independently of the mevalonate pathway. *J. Biol. Chem.* **288,** 27444–27455 (2013).
41. Moolenbeek, C. & Ruitenberg, E. J. The "Swiss roll": a simple technique for histological studies of the rodent intestine. *Lab. Anim.* **15,** 57–59 (1981).
42. Fan, T. W., Lane, A. N., Higashi, R. M. & Yan, J. Stable isotope resolved metabolomics of lung cancer in a SCID mouse model. *Metabolomics* **7,** 257–269 (2011).
43. Lane, A. N., Fan, T. W., Xie, Z., Moseley, H. N. & Higashi, R. M. Isotopomer analysis of lipid biosynthesis by high resolution mass spectrometry and NMR. *Anal. Chim. Acta* **651,** 201–208 (2009).
44. Palanker, L., Tennessen, J. M., Lam, G. & Thummel, C. S. *Drosophila* HNF4 regulates lipid mobilization and β-oxidation. *Cell Metab.* **9,** 228–239 (2009).
45. Jia, H., Liu, Y., Yan, W. & Jia, J. PP4 and PP2A regulate Hedgehog signaling by controlling Smo and Ci phosphorylation. *Development* **136,** 307–316 (2009).
46. Hipkiss, A. R. On why decreasing protein synthesis can increase lifespan. *Mech. Ageing Dev.* **128,** 412–414 (2007).
47. Stenesen, D. *et al.* Adenosine nucleotide biosynthesis and AMPK regulate adult life span and mediate the longevity benefit of caloric restriction in flies. *Cell Metab.* **17,** 101–112 (2013).
48. Alfa, R. W. *et al.* Suppression of insulin production and secretion by a decretin hormone. *Cell Metab.* **21,** 323–333 (2015).
49. Park, J. H. & Kwon, J. Y. A systematic analysis of *Drosophila* gustatory receptor gene expression in abdominal neurons which project to the central nervous system. *Mol. Cells* **32,** 375–381 (2011).
50. Scopelliti, A. *et al.* Local control of intestinal stem cell homeostasis by enteroendocrine cells in the adult *Drosophila* midgut. *Curr. Biol.* **24,** 1199–1211 (2014).
51. Wang, P. & Hou, S. X. Regulation of intestinal stem cells in mammals and *Drosophila*. *J. Cell. Physiol.* **222,** 33–37 (2010).
52. Jiang, H. & Edgar, B. A. EGFR signaling regulates the proliferation of *Drosophila* adult midgut progenitors. *Development* **136,** 483–493 (2009).
53. Gutzwiller, L. M. *et al.* Proneural and abdominal Hox inputs synergize to promote sensory organ formation in the *Drosophila* abdomen. *Dev. Biol.* **348,** 231–243 (2010).
54. Jiang, K. *et al.* Hedgehog-regulated atypical PKC promotes phosphorylation and activation of Smoothened and Cubitus interruptus in *Drosophila*. *Proc. Natl Acad. Sci. USA* **111,** E4842–E4850 (2014).
55. Liu, Y., Cao, X., Jiang, J. & Jia, J. Fused–Costal2 protein complex regulates Hedgehog-induced Smo phosphorylation and cell-surface accumulation. *Genes Dev.* **21,** 1949–1963 (2007).
56. Yang, L., Gal, J., Chen, J. & Zhu, H. Self-assembled FUS binds active chromatin and regulates gene transcription. *Proc. Natl Acad. Sci. USA* **111,** 17809–17814 (2014).
57. Reiher, W. *et al.* Peptidomics and peptide hormone processing in the *Drosophila* midgut. *J. Proteome Res.* **10,** 1881–1892 (2011).
58. Bergland, G. Minisymposium: The Malmö Diet and Cancer Study. Design, biological bank and biomarker programme. *J. Intern. Med.* **233,** 39–79 (1993).
59. Persson, M., Berglund, G., Nelson, J. J. & Hedblad, B. Lp-PLA2 activity and mass are associated with increased incidence of ischemic stroke: a population-based cohort study from Malmö, Sweden. *Atherosclerosis* **200,** 191–198 (2008).
60. Matthews, D. R. *et al.* Homeostasis model assessment: insulin resistance and β-cell function from fasting plasma glucose and insulin concentrations in man. *Diabetologia* **28,** 412–419 (1985).
61. Alberti, K. G., Zimmet, P. & Shaw, J. Metabolic syndrome—a new world-wide definition. A Consensus Statement from the International Diabetes Federation. *Diabet. Med.* **23,** 469–480 (2006).
62. Ernst, A., Hellmich, S. & Bergmann, A. Proneurotensin 1–117, a stable neurotensin precursor fragment identified in human circulation. *Peptides* **27,** 1787–1793 (2006).
63. Enhörning, S. *et al.* Plasma copeptin and the risk of diabetes mellitus. *Circulation* **121,** 2102–2108 (2010).
64. Li, J. *et al.* PI3K p110α/Akt signaling negatively regulates secretion of the intestinal peptide neurotensin through interference of granule transport. *Mol. Endocrinol.* **26,** 1380–1393 (2012).

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 1 | NT deficiency does not affect body length or small intestine morphology. a, Body length did not differ significantly between genotypes for either male ($Nt^{+/+}$ $n=13$, $Nt^{-/-}$ $n=12$) or female ($Nt^{+/+}$ $n=12$, $Nt^{-/-}$ $n=12$) mice. **b**–**f**, The average weight (**b**) and length (**c**) of the small intestine was similar between the genotypes ($n=7$). Proximal intestinal samples were H&E stained (**d**; scale bar, 50 µm); villus height (**e**; $n=6$) and crypt numbers (**f**; $n=6$) have no significant differences between genotypes. Mice (7 months old) for all experiments were maintained on standard chow. All data are mean ± s.d. Two-sided Student's *t*-test for all figures.

© 2016 Macmillan Publishers Limited. All rights reserved

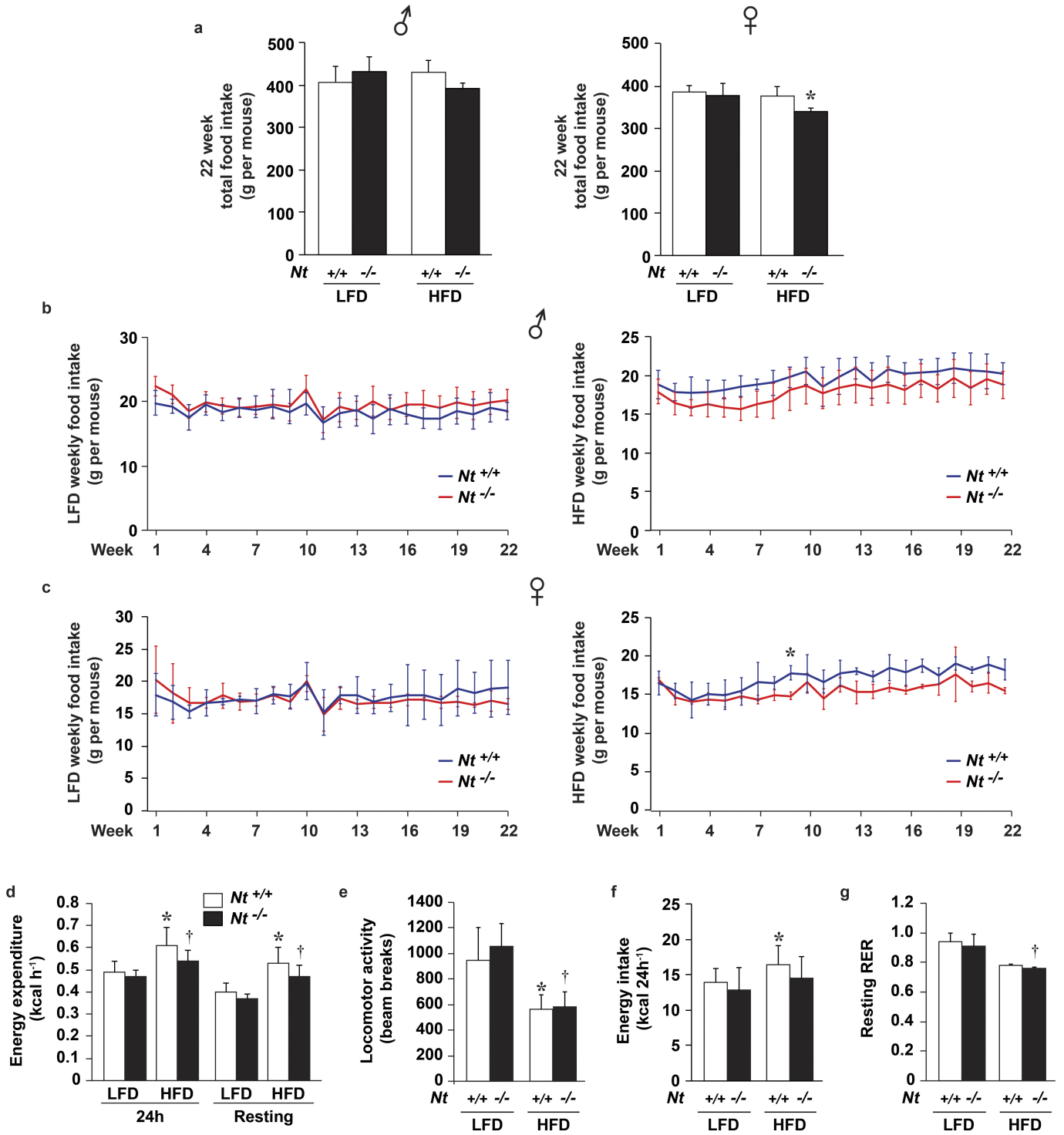


Extended Data Figure 2 | See next page for caption.

© 2016 Macmillan Publishers Limited. All rights reserved

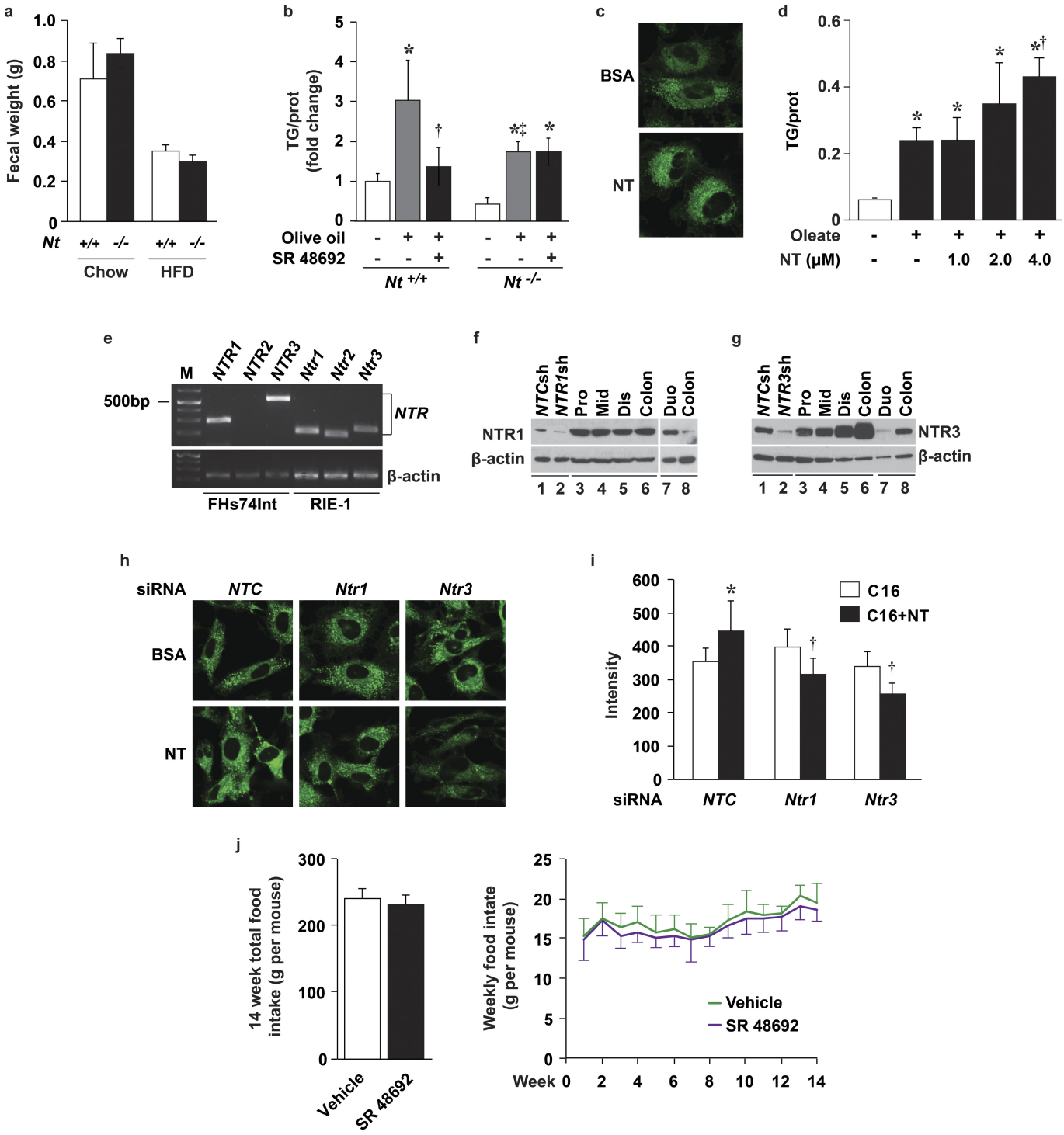
Extended Data Figure 2 | NT deficiency inhibits adiposity, inflammation and improves insulin/Akt signalling. a–**c**, Epididymal fat pad before (**a**; arrows) and after (**b**) dissection, and retroperitoneal fat pad (**c**; arrows) from male mice maintained on a normal chow ($n = 5$). **d**–**f**, Representative epididymal (**d**; arrows), retroperitoneal (**e**) and pericardial fat pads (**f**) of male mice fed a HFD for 24 weeks ($n = 5$). **g**, Weekly body weight (male $Nt^{+/+}$ and $Nt^{-/-}$ $n = 13$) and (female $Nt^{+/+}$ $n = 12$, $Nt^{-/-}$ $n = 14$) of mice on a LFD for 22 weeks. $^{*}P < 0.05$ $Nt^{+/+}$ versus $Nt^{-/-}$ in male mice. **h**, Fat composition in male mice fed a LFD or HFD for 24 weeks ($n = 5$). $^{*}P < 0.05$ versus LFD fat; $^{\dagger}P < 0.05$ versus $Nt^{+/+}$ LFD fat; $^{\ddagger}P < 0.05$ versus $Nt^{+/+}$ HFD fat; $^{\S}P < 0.05$ versus $Nt^{+/+}$ HFD lean. **i**, $Nt^{+/+}$ and $Nt^{-/-}$ mice fed a LFD or HFD for 22 weeks after weaning were fasted overnight. Saline or insulin (5 units) was injected into the inferior vena cava (IVC). Five minutes after the injection, liver tissues were collected and protein extracts analysed by western blot. A representative result is shown from three separate experiments. **j**, Male $Nt^{+/+}$ and $Nt^{-/-}$ mice (7 months old) maintained on standard chow were fasted overnight followed by glucose (2 mg per kg body weight) administration by gavage ($n = 5$). Blood samples were collected both before and at the indicated times after glucose administration for measurement of plasma insulin (left) and glucose (right) levels. No statistically significant differences in either insulin or glucose levels were apparent between genotypes. **k**, Male $Nt^{+/+}$ and $Nt^{-/-}$ mice (12 months old) fed a standard diet after weaning were fasted for 16 h or fasted for 16 h and refed for 4 h ($n = 3$). Blood was collected from the IVC and plasma was used to measure insulin. $^{*}P < 0.05$ versus fasted $Nt^{+/+}$ mice. **l**, H&E staining of epididymal fat pad from male mice shown in Fig. 1g demonstrating inflammatory cells (arrows) (scale bar, 100 μm). **m**, F4/80-positive macrophages (arrows) in epididymal fat pad from HFD-fed mice (scale bar, 100 μm) ($n = 5$). All data are mean ± s.d. Linear mixed model for **g**; analysis of covariance (ANCOVA) with Holm's *P*-value adjustment for **h**; ANOVA with Holm's *P*-value adjustment for **j**, **k**. (See Supplementary Fig. 1 for gel source data.)

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 3 | Food intake and indirect calorimetry measurements. **a**–**c**, Food intake was measured weekly for $Nt^{+/+}$ and $Nt^{-/-}$ male (LFD $n = 3$ and HFD $n = 4$) and female (LFD $n = 3$ and HFD $n = 4$) mice. Analysis of cumulative food intake for 22 weeks did not show a significant difference except in female mice fed a HFD. $^{*}P < 0.05$ versus female $Nt^{+/+}$ fed a HFD (**a**). For weekly food intake, there was no significant difference between genotypes in male mice on both diets (**b**) and female mice fed a LFD (**c**, left); food intake on week 9 in female mice fed a HFD reached significance between genotypes. $^{*}P < 0.05$ versus $Nt^{-/-}$ (**c**, right). **d**–**g**, Male mice on LFD and HFD diets for 24 weeks were placed in individual cages and indirect calorimetric analysis was performed ($n = 10$). Energy expenditure is presented by the average kcal h^{-1} in 24 h and in resting period (**d**). Locomotor activity represents counts of beam breaks in a 30-min period (**e**). Energy intake represents the food intake in kcal for 24 h (**f**). Resting RER is presented in **g**. $^{*}P < 0.05$ versus LFD in $Nt^{+/+}$ mice; $^{\dagger}P < 0.05$ versus LFD in $Nt^{-/-}$ mice. All data are mean ± s.d. Two-sided Student's *t*-test per diet and sex groups for **a**; linear mixed model for **b**, **c**; ANCOVA for **d**–**g**.

© 2016 Macmillan Publishers Limited. All rights reserved

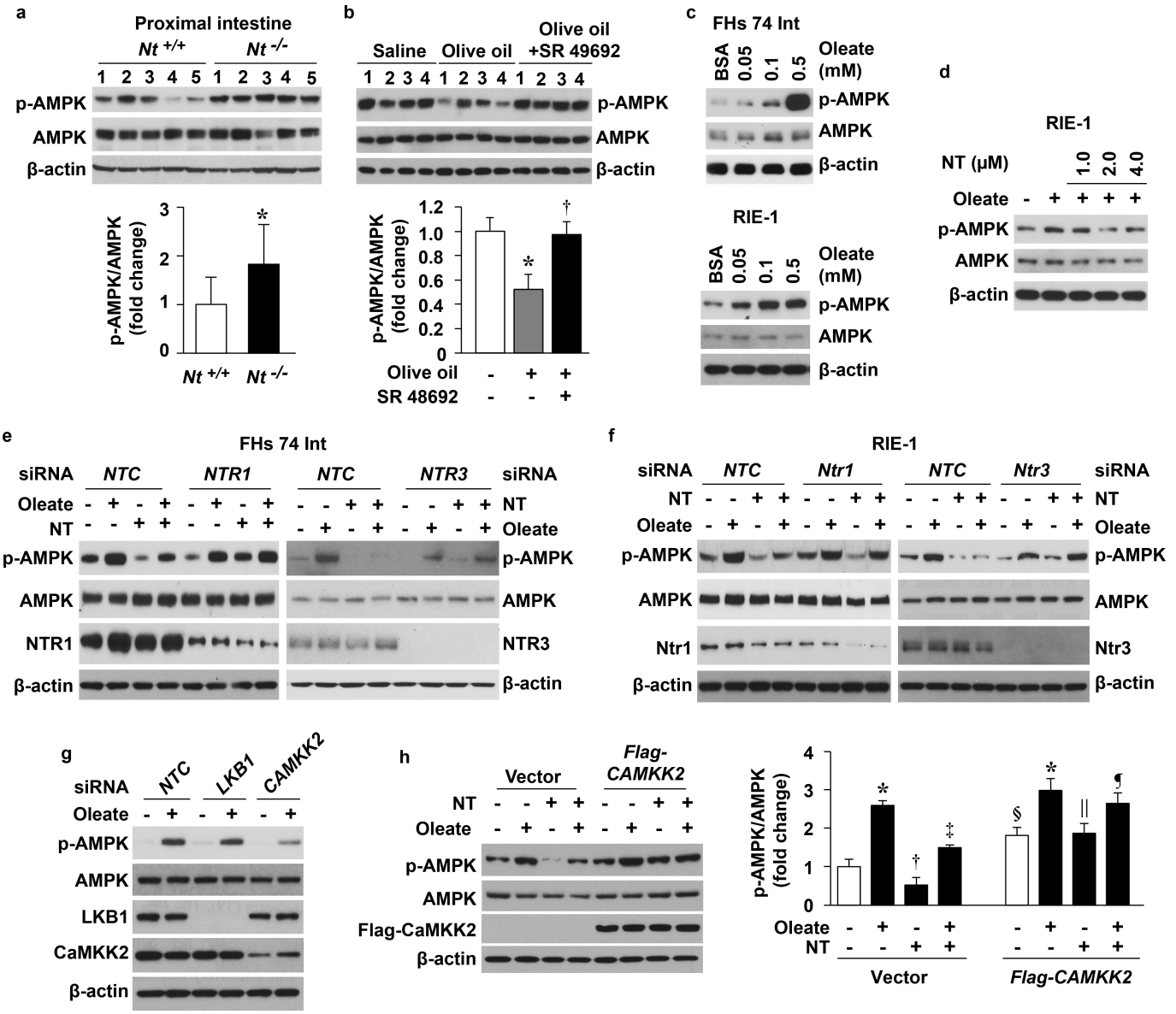


Extended Data Figure 4 | See next page for caption.

© 2016 Macmillan Publishers Limited. All rights reserved

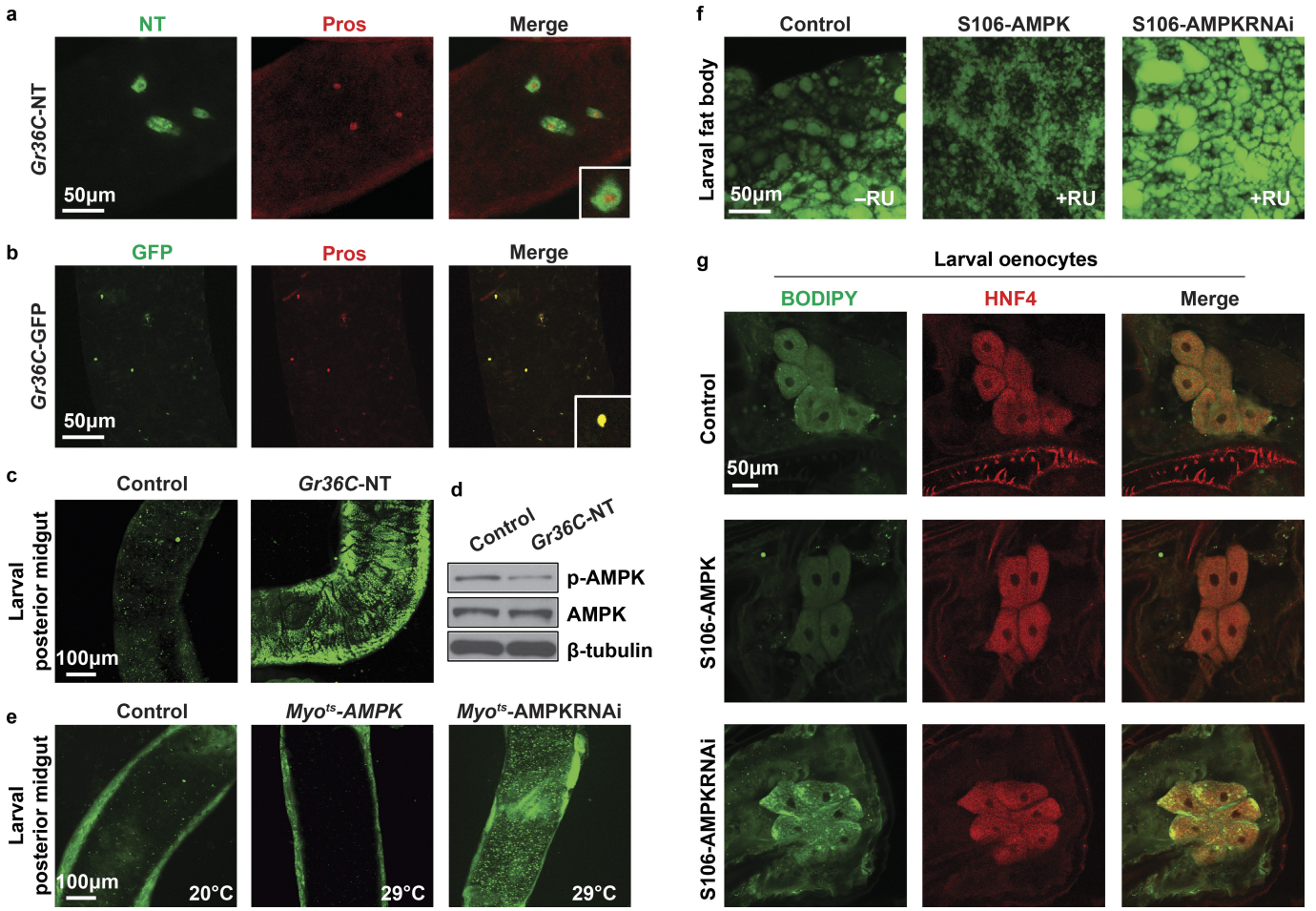
Extended Data Figure 4 | NT promotes intestinal cell lipid absorption or accumulation. a, Faecal weight of male $Nt^{+/+}$ and $Nt^{-/-}$ mice fed either normal chow (NC; $Nt^{+/+}$ $n=4$, $Nt^{-/-}$ $n=5$) or a HFD ($Nt^{+/+}$ $n=7$, $Nt^{-/-}$ $n=9$) for 24 weeks; there was no significant difference between genotypes. **b**, Male $Nt^{+/+}$ and $Nt^{-/-}$ mice ($n=4$) on normal chow were fasted overnight. Mice were injected with either saline or SR 48692 (2.5 mg per kg body weight, i.p). Thirty minutes after the injection, mice were given olive oil (10 μl per g body weight) by oral gavage and then killed. The proximal small intestine was excised and intestinal TG content measured. $*P<0.05$ versus saline in $Nt^{+/+}$ and $Nt^{-/-}$ mice; $\dagger P<0.05$ versus olive oil only in $Nt^{+/+}$ mice; $\ddagger P<0.05$ versus olive oil only in $Nt^{+/+}$ mice. **c**, RIE-1 cells were pre-treated with BSA or NT (2 μM) for 30 min, followed by incubation with BSA-conjugated BODIPY FL C_{16} for 15 min, and labelled lipids were visualized by confocal microscopy (original magnification, ×180). Representative images at ×60, zoom 3 are from three experiments. **d**, RIE-1 cells were pre-treated with or without NT at different concentrations for 30 min followed by the addition of BSA-conjugated oleate (0.1 mM) and further incubation overnight. Cells were collected, lysed and TG was measured ($n=3$). $*P<0.05$ versus BSA only; $\dagger P<0.05$ versus oleate only. **e**, Total RNA was isolated from human (FHs 74 Int) and rat (RIE-1) small intestinal epithelial cells and reverse transcription polymerase chain reaction (RT–PCR) was performed using specific primers targeting human or rat *NTR1*, *2* and *3*. **f**, **g**, Expression of NTR1 (**f**) and NTR3 (**g**) was analysed by western blot from mucosa scraped from mouse proximal (pro), middle (mid), distal (dis) small intestine and colon (lanes 3–6) as well as human duodenum (duo) and colon (lanes 7–8). Proteins from HepG2 (human hepatocellular carcinoma cell line) (**f**; lanes 1–2) and Caco-2 (human colon cancer cell line) (**g**; lanes 1–2) cells stably expressing *NTR1*, *NTR3*, or control (*NTC*) shRNA were used as positive and negative controls. **h**, **i**, RIE-1 cells transfected with non-targeting control (*NTC*) siRNA or *Ntr1* or *Ntr3* siRNA for 72 h were treated with or without NT (2 μM) for 30 min followed by BSA-conjugated BODIPY FL C_{16} (C_{16}) for 15 min, and imaged by confocal microscopy (original magnification, ×180) to quantify fluorescence (**h**) and intensity (**i**) as described in Methods ($n=30$ cells). $*P<0.05$ versus C_{16} in *NTC* siRNA; $\dagger P<0.05$ versus C_{16} plus NT in *NTC* siRNA. **j**, Cumulative (left) and weekly (right) food intake was measured in male wild-type C57BL/6 mice fed a HFD and chronically treated with either SR 48692 ($n=13$) or vehicle ($n=12$). Neither analysis demonstrated a significant difference. All data are mean ± s.d. ANOVA with Holm's *P*-value adjustment for **a**, **b**, **d**, **i**; two-sided Student's *t*-test for **j** (left); linear mixed model for **j** (right). (See Supplementary Fig. 1 for gel source data.)

© 2016 Macmillan Publishers Limited. All rights reserved



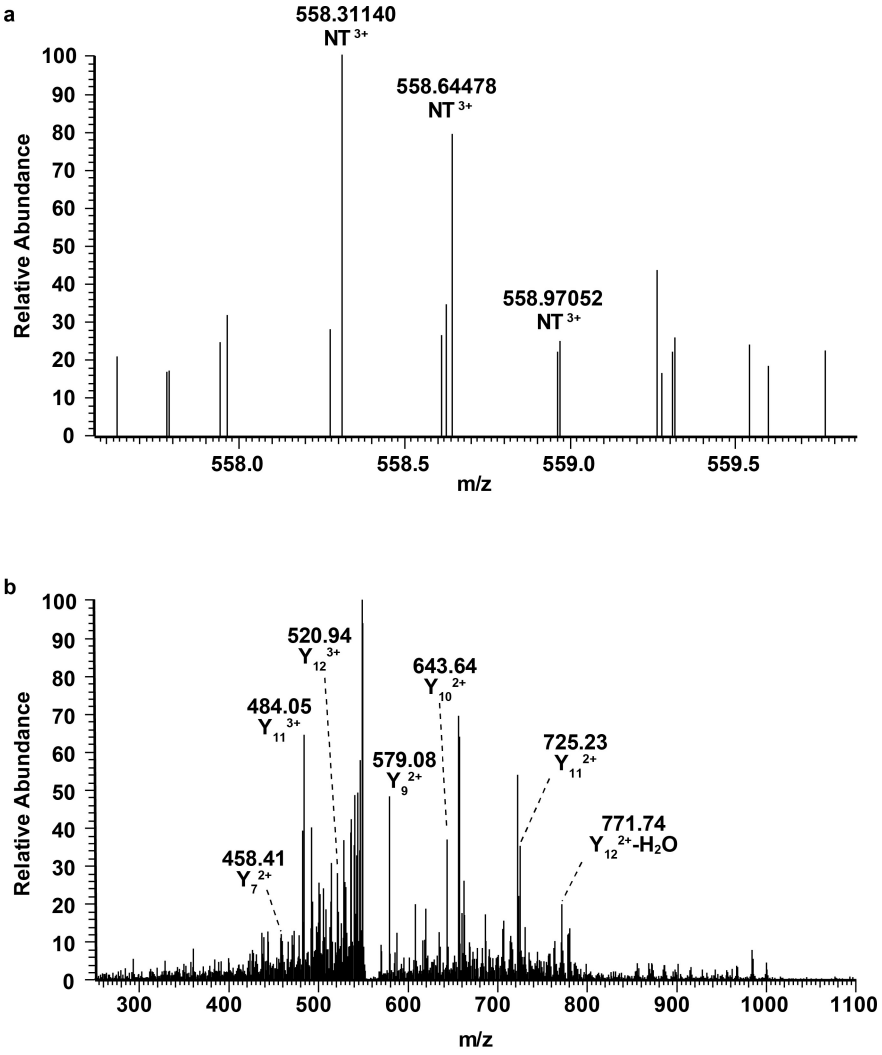
Extended Data Figure 5 | NT negatively regulates AMPK activity. **a**, Western blotting and densitometry of p-AMPK in proximal intestinal mucosa of male mice (12 months old) maintained on standard chow and fasted for 24 h ($n = 9$). *$P < 0.05$ versus $Nt^{+/+}$. **b**, Western blotting and densitometry of p-AMPK levels using samples described in Fig. 2e (n = 8). *$P < 0.05$ versus saline; †$P < 0.05$ versus olive oil alone. **c**, FHs 74 Int (top) and RIE-1 (bottom) cells were treated with the indicated concentrations of oleate for 1 h and lysates were analysed by western blotting. **d**, RIE-1 cells were pre-treated with or without NT for 30 min followed by combined treatment with BSA or oleate (0.1 mM) for 1 h and western blot analysis. **e**, FHs 74 Int cells transfected with either human *NTR1* or *NTC* siRNA (100 nM)), or either human *NTR3* or *NTC* siRNA (20 nM) as indicated for 72 h were pre-treated with or without NT (2 μM) for 30 min followed by treatment with oleate (0.1 mM) or BSA for 1 h and western blot analysis. **f**, RIE-1 cells were transfected with rat siRNAs and treated with NT and oleate as described in **e**. **g**, FHs 74 Int cells transfected with *LKB1*, *CAMKK2* or control (all 40 nM) siRNAs for 3 days were pre-treated with NT (2 μM) for 30 min followed by oleate (0.1 mM) for 1 h and cell extracts were analysed by western blotting. **h**, FHs 74 Int cells transfected with *Flag-CAMKK2* and control vector for 48 h were pre-treated with or without NT (2 μM) followed by oleate (0.1 mM) or BSA for 1 h and analysed by western blotting (left); p-AMPK levels were determined as in **b** from three separate experiments (right). *$P < 0.05$ versus BSA in vector- and *CAMKK2*-transfected cells, respectively; †$P < 0.05$ versus BSA in vector-transfected cells; ‡$P < 0.05$ versus oleate in vector-transfected cells; §$P < 0.05$ versus BSA in vector-transfected cells; ||$P < 0.05$ versus NT alone in vector-transfected cells; ¶$P < 0.05$ versus oleate plus NT in vector-transfected cells. All data are mean ± s.d. Two-sided Student's *t*-test for **a** (bottom); ANOVA with Holm's *P*-value adjustment for **b** (bottom), **h** (right) (See Supplementary Fig. 1 for gel source data.)

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 6 | NT regulates lipid droplet accumulation and AMPK activation in *Drosophila* midgut, fat body, and oenocytes. **a**, Midgut from *Gr36C*-NT adult (7 days old) was stained for NT and Prospero (Pros). Insets in **a** and **b** are high-magnification images of individual Pros-expressing enteroendocrine cells, co-expressing either NT or nuclear green fluorescent protein (GFP) (see later), respectively ($n = 3$). Scale bar, 50 μm. **b**, Similar experiment using *Gr36C*-Gal4 to drive nuclear GFP expression, demonstrating co-localization of GFP and Pros in nuclei of enteroendocrine cells in adult flies (7 days old) ($n = 3$). Scale bar, 50 μm. **c**, NT expression promotes the accumulation of lipid droplets. Midguts from either control *Gr36C*-Gal4 (left; 100%, $n = 15$) or *Gr36C*-NT (right; 100%, $n = 19$) 3rd instar larvae stained with BODIPY. Similar results were obtained with Nile Red staining (data not shown). Scale bar, 100 μm. **d**, Western blot analysis of p-AMPK and AMPK levels in gastrointestinal tract of *Gr36C*-Gal4 control and *Gr36C*-NT 3rd instar larvae. **e**, Conditional expression of AMPK (middle; *Myo*^{ts}-AMPK, 29 °C, 86%, $n = 14$) or AMPK RNAi (right; *Myo*^{ts}-AMPKRNAi^{25931}, 29 °C, 91%, $n = 11$) leads to either suppression or enhancement of lipid accumulation (visualized with BODIPY) as compared to control (left; *Myo*^{ts}-AMPK, 20 °C, 100%, $n = 9$) 3rd instar larvae. Embryos were raised at 20 °C and 96 h after egg laying were switched to the non-permissive temperature (29 °C) to induce Gal4 expression. AMPK overexpression and RNAi inhibition were monitored by western blotting with AMPK antibody (data not shown). Scale bar, 100 μm. **f**, Similar results were obtained in 3rd instar larval fat bodies using a RU486- inducible *S106*-Gal4 driver to drive AMPK without RU486 (left; control, 100%, $n = 9$), AMPK with RU486 (middle; 89%, $n = 9$) or AMPK RNAi with RU486 (right; 100%, $n = 10$) expression. Scale bar, 50 μm. **g**, Lipid accumulation in oenocytes (HNF-positive) visualized with BODIPY as in **e**. Genotypes and fluorescent stains are indicated. Compare middle panels (100%, $n = 6$) and bottom panels (89%, $n = 9$) to top control panels (100%, $n = 8$). Scale bar, 50 μm. (See Supplementary Fig. 1 for gel source data.)

© 2016 Macmillan Publishers Limited. All rights reserved

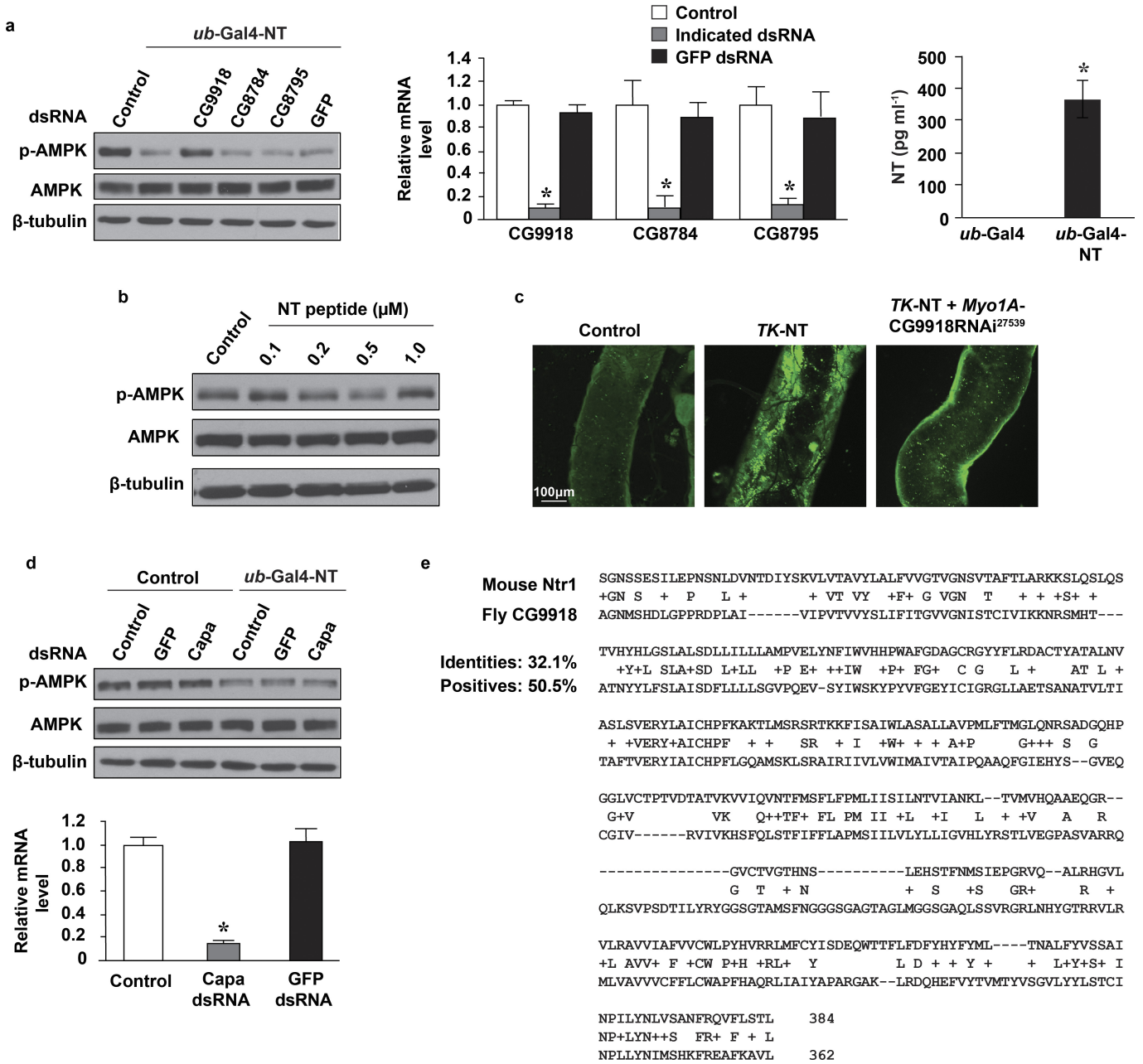


c

NT(1-13) concentration in S2 cell culture medium	67 x 10^{-12} g/ml
NT(1-13) in control fly gut	N.D.
Average amount of NT(1-13) in Gr36C-NT fly gut	3.1 x 10^{-15} g/gut

Extended Data Figure 7 | LC–MS/MS analysis of the processed NT peptide. a, The mass spectrum of the triply charged NT (NT^{3+}) peptide eluted at 18.96 min in conditioned medium of S2 cells expressing full-length human NT precursor. The labelled three peaks are the isotopic envelope of NT^{3+} with a mass accuracy less than 2 p.p.m. from the theoretical *m*/*z* value of 558.31050. **b**, The tandem mass spectrum of the triply charged NT^{3+} peptide. The *m*/*z* values of major fragment ions are designated, confirming the peptide as biologically active 13-amino-acid NT. **c**, The amount of NT in S2 medium and fly gastrointestinal tracts from adult *Gr36C*-NT flies (350 guts collected for each genotype) was quantified. N.D., not detected.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 8 | CG9918 (mouse NTR1 analogue) RNAi blocks the NT-mediated decrease of p-AMPK levels in *Drosophila* S2 cells. **a**, S2 cells were transfected with *ub*-Gal4 plus UAST-NT (*ub*-Gal4-NT) or control (*ub*-Gal4) vector and treated with the indicated double-stranded RNAs (dsRNAs) to knockdown individual receptors. Cell lysates were subjected to western blot with the indicated antibodies to examine the activation of AMPK (left). The efficiency of RNAi knockdown was monitored by real-time PCR (middle; $n = 3$); $^*P < 0.05$ versus control dsRNA. Medium from cells expressing either *ub*-Gal4 alone or *ub*-Gal4-NT was collected to examine NT levels by enzyme immunoassay (EIA)[36,64] (right; $n = 6$); $^*P < 0.05$ versus *ub*-Gal4. **b**, S2 cells were treated with the indicated concentrations of NT peptide for 1 h and AMPK activation was monitored as in **a**. NT treatment (0.2, 0.5 μM) decreased p-AMPK to levels similar to those observed in *ub*-Gal4-NT-transfected cells where NT levels reach approximately the same concentration (~350 pg ml^{-1}, 0.2 μM). **c**, Midguts of 3rd instar larvae from the indicated genotypes were stained with BODIPY to monitor the accumulation of lipid droplets. Left, *w1118* control midgut accumulates low level of lipid (100%, $n = 8$). Middle, larval midgut constitutively expressing NT by the TK promoter (TK-NT) accumulates high levels of lipid (100%, $n = 8$). Right, larval midgut co-expressing TK-NT and *Myo1A*-CG9918RNAi[27539] accumulated much lower levels of lipid (87%, $n = 15$) compared to TK-NT midgut (middle). Scale bar, 100 μm. **d**, S2 cells were transfected with *ub*-Gal4-NT or control vector, treated with the indicated dsRNAs, and cell lysates were analysed as in **a**. Inactivation of Capability (Capa; CG15520), the Pyrokinin-1 in *Drosophila*, does not alter the levels of p-AMPK in either the presence or absence of NT (top), indicating that NT prevents AMPK activation independently of Pyrokinin-1. Capa RNAi efficiency was monitored by real time-PCR (bottom; $n = 3$). $^*P < 0.05$ versus control dsRNA. **e**, Amino acid sequence alignment of mouse NTR1 and *Drosophila* CG9918, identities and conserved residues (+) are indicated. All data are mean ± s.d. ANOVA for **a** (middle), **d** (bottom); two-sided Wilcoxon rank-sum test for **a** (right). (See Supplementary Fig. 1 for gel source data.)

© 2016 Macmillan Publishers Limited. All rights reserved

Extended Data Table 1 | Clinical characteristics of the MDC-CC

Characteristic	Value	N
Age (years)	57.7 ± 6.0	4,626
Female sex, n (%)	2661 (57.5)	4,626
Body Mass Index (kg/m^2)	25.8 ± 3.9	4,626
Waist circumference (cm)	84.0 ± 12.9	4,625
Fasting blood glucose (mM)	5.2 ± 1.4	4,468
Fasting insulin concentration (mU/L)	7.0 (4.0-9.0)	4,468
HOMA-IR	1.5 (0.9-2.2)	4,468

Data are given as mean ± s.d. for normally distributed variables, and as median and interquartile range for fasting insulin concentration. Categorical data are presented as numbers (percentages). *N* denotes the number with complete data, thus, included in analyses.

© 2016 Macmillan Publishers Limited. All rights reserved

Extended Data Table 2 | Continuous values of pro-NT and association with continuous values of clinical outcomes

Metabolic trait	β-coefficient ± SE (per SD increment of log-transformed pro-NT)	p-value
Body Mass Index (kg/m^2)	0.14 ± 0.06	0.016
Waist circumference (cm)	0.33 ± 0.15	0.026
HOMA-IR (mmol/L*mU/L)	0.13 ± 0.015	<0.001

Based on linear regression model of log-transformed fasting plasma concentration of pro-NT expressed as per 1 s.d. increment in log-transformed value. Standard error (s.e.) of beta coefficient.

© 2016 Macmillan Publishers Limited. All rights reserved

LETTER

doi:10.1038/nature17989

Maintenance and propagation of a deleterious mitochondrial genome by the mitochondrial unfolded protein response

Yi-Fan Lin[1], Anna M. Schulz[1], Mark W. Pellegrino[1], Yun Lu[2], Shai Shaham[2] & Cole M. Haynes[1,3]

Mitochondrial genomes (mitochondrial DNA, mtDNA) encode essential oxidative phosphorylation (OXPHOS) components. Because hundreds of mtDNAs exist per cell, a deletion in a single mtDNA has little impact. However, if the deletion genome is enriched, OXPHOS declines, resulting in cellular dysfunction. For example, Kearns–Sayre syndrome is caused by a single heteroplasmic mtDNA deletion. More broadly, mtDNA deletion accumulation has been observed in individual muscle cells[1] and dopaminergic neurons[2] during ageing. It is unclear how mtDNA deletions are tolerated or how they are propagated in somatic cells. One mechanism by which cells respond to OXPHOS dysfunction is by activating the mitochondrial unfolded protein response (UPR^{mt}), a transcriptional response mediated by the transcription factor ATFS-1 that promotes the recovery and regeneration of defective mitochondria[3,4]. Here we investigate the role of ATFS-1 in the maintenance and propagation of a deleterious mtDNA in a heteroplasmic *Caenorhabditis elegans* strain that stably expresses wild-type mtDNA and mtDNA with a 3.1-kilobase deletion (ΔmtDNA) lacking four essential genes[5]. The heteroplasmic strain, which has 60% ΔmtDNA, displays modest mitochondrial dysfunction and constitutive UPR^{mt} activation. ATFS-1 impairment reduced the ΔmtDNA nearly tenfold, decreasing the total percentage to 7%. We propose that in the context of mtDNA heteroplasmy, UPR^{mt} activation caused by OXPHOS defects propagates or maintains the deleterious mtDNA in an attempt to recover OXPHOS activity by promoting mitochondrial biogenesis and dynamics.

mtDNA mutations and deletions are relatively common within the mtDNA population of ageing post-mitotic cells but typically represent a minor percentage of the total mtDNA[6,7]. However, if the deleterious mtDNA reaches a high enough percentage relative to wild-type mtDNA, it can, depending on the severity of the mtDNA lesion, become toxic and lead to disease[8]. As mtDNA encodes 13 (12 in *C. elegans*) essential components of the respiratory chain and ATP synthase, along with rRNAs and tRNAs required for their synthesis, an increase in deleterious heteroplasmy can perturb multiple cellular activities reliant on OXPHOS. However, the mechanism(s) that promote tolerance to large mtDNA deletions or contribute to their propagation are unknown.

Here we examine the role of the UPR^{mt} on the maintenance and propagation of a mtDNA with a 3.1 kb deletion (ΔmtDNA) encoding four essential OXPHOS genes (Fig. 1a) using a heteroplasmic *C. elegans* strain[5]. The UPR^{mt} is regulated by the transcription factor ATFS-1, which is normally efficiently imported into mitochondria and degraded. However, during mitochondrial stress or OXPHOS dysfunction, a percentage of ATFS-1 fails to be imported and traffics to the nucleus to activate a broad transcriptional program (over 500 transcripts) that promotes repair and recovery of mitochondrial function[3,4].

To examine the effect of ΔmtDNA on UPR^{mt} activation, ΔmtDNA was crossed into a transcriptional reporter worm harbouring the *hsp-6$_{pr}$::gfp* transgene used to monitor UPR^{mt} activation[9]. Consistent with previous studies demonstrating that perturbation of nuclear-encoded OXPHOS components activated the UPR^{mt} (Extended Data Fig. 1a)[4,10], ΔmtDNA also modestly activated the UPR^{mt}, which required *atfs-1* (Fig. 1b). Consistent with a previous report[5], ΔmtDNA made up 60% of all mtDNAs (Fig. 1c). And, ΔmtDNA caused a significant reduction in basal oxygen consumption as well as total respiratory capacity (Fig. 1d), suggesting 60% ΔmtDNA perturbs OXPHOS and activates the UPR^{mt}.

Surprisingly, the development of worms harbouring ΔmtDNA was unaffected by RNAi-mediated knockdown of *atfs-1* (*atfs-1*(RNAi)), in stark contrast to worms with hypomorphic mutations in nuclear-encoded OXPHOS components (Fig. 1e)[4,11], indicating the UPR^{mt} is not required for development when the OXPHOS defect derives from a mtDNA deletion. Thus, we examined the effect of UPR^{mt} inhibition on ΔmtDNA levels by impairing several components required for *hsp-6$_{pr}$::gfp* induction[4,12]. Notably, *atfs-1* deletion or knockdown caused a dramatic reduction in ΔmtDNA levels, shifting the percentage from 60% to 7% (Fig. 1c and Extended Data Fig. 1b–d), probably explaining the normal growth rate (Fig. 1e). ΔmtDNA quantification in individual wild-type or *atfs-1*-deletion worms was consistent with that observed in larger worm populations and in some cases ΔmtDNA was reduced below the limit of detection in worms lacking ATFS-1 (Extended Data Fig. 1e). Therefore, ATFS-1 and UPR^{mt} activation are required to maintain the deleterious mtDNA.

A mechanism by which deleterious mtDNA levels can be altered involves the germ-line bottleneck where only a small number of mtDNAs are passed maternally to the next generation, allowing shifts in heteroplasmy[8]. Because the *atfs-1*-deletion strain was generated via mating, it is unclear if the shift in ΔmtDNA occurred during germ-line transmission, somatic cell division and growth, or both. To examine the role of ATFS-1 in ΔmtDNA maintenance specifically in somatic cells, we employed *glp-4(bn2)* worms that lack germlines when raised at the restrictive temperature. Interestingly, exposure of worms to *atfs-1* RNAi from the L1 stage to adulthood also depleted ΔmtDNA relative to the identical worm population raised on control RNAi (Fig. 1f), consistent with a role for the UPR^{mt} in maintaining ΔmtDNA levels in post-mitotic somatic cells.

Because mitochondrial autophagy (mitophagy) could potentially eliminate ΔmtDNAs when *atfs-1* is inhibited, we examined the interaction between ΔmtDNA, *atfs-1* and known mitophagy components. Mitophagy involves the recognition of defective mitochondria by the kinase PINK-1. Once PINK-1 accumulates, it recruits the ubiquitin ligase Parkin (PDR-1 in *C. elegans*) to the mitochondrial outer membrane, which directs the damaged organelle to lysosomes for degradation[13]. As described previously, *pdr-1* deletion resulted in increased

[1]Cell Biology Program, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA. [2]Laboratory of Developmental Genetics, The Rockefeller University, New York, New York 10065, USA. [3]BCMB Allied Program, Weill Cornell Medical College, 1300 York Avenue, New York, New York 10065, USA.

© 2016 Macmillan Publishers Limited. All rights reserved

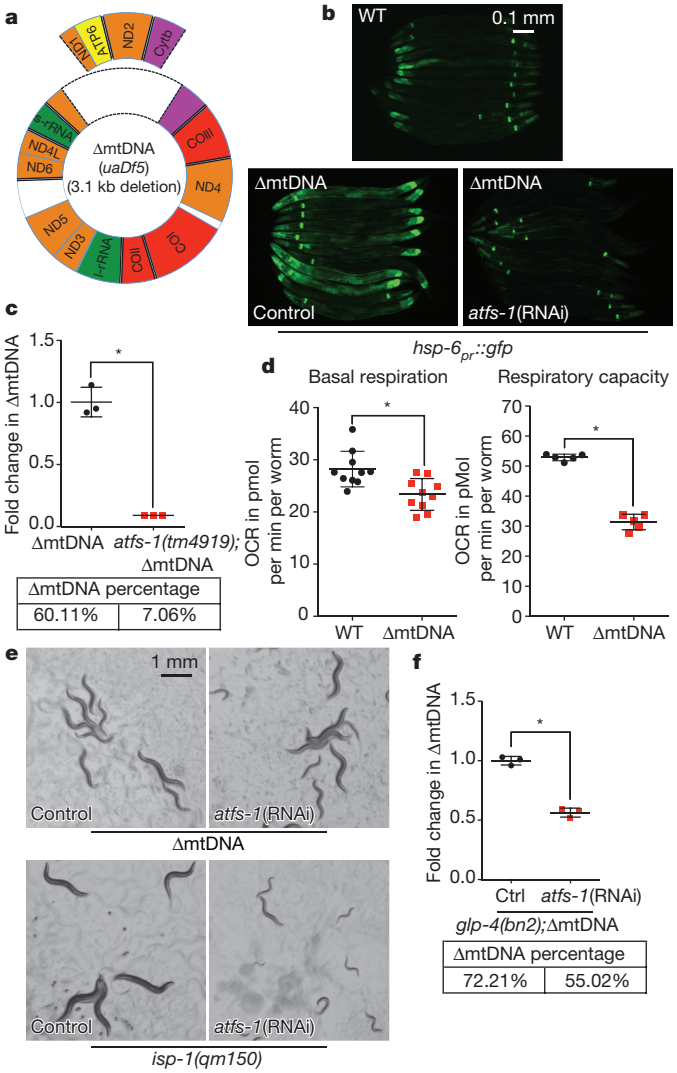


Figure 1 | ATFS-1 is required to maintain a deleterious mtDNA. **a**, Comparison of wild-type and *uaDf5* deletion (ΔmtDNA) mtDNAs. **b**, *hsp-6*$_{pr}$*::gfp* in wild-type or ΔmtDNA worms on control or *atfs-1* RNAi. Scale bar, 0.1 mm. **c**, ΔmtDNA quantification as determined by qPCR in ΔmtDNA or *atfs-1(tm4919);*ΔmtDNA worms. $n = 3$; error bars, mean ± s.d.; $^*P < 0.03$ (Student's *t*-test). **d**, Oxygen consumption rates (OCR) in wild-type and ΔmtDNA worms. $n = 5$ and $n = 10$, respectively; error bars, mean ± s.d.; $^*P < 0.03$ (Student's *t*-test). **e**, ΔmtDNA or *isp-1(qm150)* worms raised on control or *atfs-1* RNAi. Scale bar, 1 mm. **f**, ΔmtDNA quantification as determined by qPCR in *glp-4(bn2);*ΔmtDNA worms on control or *atfs-1* RNAi. $n = 3$; error bars, mean ± s.d.; $^*P < 0.03$ (Student's *t*-test).

ΔmtDNA (Fig. 2a)[14], consistent with mitophagy targeting defective mitochondria containing relatively high levels of deleterious mtDNAs[15]. However, the reduction of ΔmtDNA caused by *atfs-1* inhibition was only partially blocked by *pink-1;pdr-1* deletion or *atg-18* deletion, which impairs general autophagy (Fig. 2a and Extended Data Fig. 2a). Additionally, *atfs-1* inhibition did not impair development of *pink-1;pdr-1*-deficient ΔmtDNA worms (Fig. 2b) despite the increased UPR^{mt} activation (Fig. 2c). As PDR-1 inhibition did not completely restore ΔmtDNA in *atfs-1*-deletion worms, ATFS-1 probably promotes ΔmtDNA propagation independent of mitophagy or other Parkin-mediated activities[16].

As modest UPR^{mt} activation was required to maintain the deletion genome (Fig. 1), we examined the effects of stronger UPR^{mt} activation on ΔmtDNA expansion or propagation. To further activate the UPR^{mt} in somatic cells, the mitochondrial protease SPG-7 was impaired by RNAi[3,4] during development in worms lacking germlines. Despite strong UPR^{mt} activation, ΔmtDNA levels did not increase (Fig. 2d and Extended Data Fig. 2b). However, because PINK-1 and Parkin are also activated by mitochondrial unfolded protein stress[17], we wondered whether PINK-1 and PDR-1 activities limit ΔmtDNA accumulation. Remarkably, ΔmtDNA levels increased threefold in *pink-1;pdr-1*-deletion worms raised on *spg-7* RNAi (Fig. 2d) suggesting that UPR^{mt} activation can promote propagation of deleterious mtDNAs, which is antagonized by mitophagy during strong mitochondrial stress. However, because wild-type mtDNAs were also increased, the ΔmtDNA percentage was unaffected (Fig. 2d). As both wild-type and ΔmtDNA were increased by *spg*-7(RNAi), ATFS-1 may promote a mitochondrial biogenesis program in response to mitochondrial dysfunction. Notably, mitochondrial mass[18] also increased in *pink-1; pdr-1*-deficient worms treated with *spg-7* RNAi (Fig. 2e) consistent with ATFS-1 mediating a compensatory mitochondrial biogenesis program that maintains ΔmtDNA. Combined, these results suggest that during strong mitochondrial stress, balanced PINK-1/PDR-1 and ATFS-1 activity limits ΔmtDNA accumulation.

To determine if ATFS-1 activation is sufficient to increase mitochondrial biogenesis and ΔmtDNAs independent of mitophagy, we used a mutant strain with constitutive UPR^{mt} activation owing to an amino acid substitution within the ATFS-1 mitochondrial targeting sequence[19]. Consistent with the UPR^{mt} promoting a mitochondrial biogenesis program, *atfs-1(et18)* animals also displayed a marked increase in the number of mitochondria (Fig. 3a). *atfs-1(et18)* worms also had increased ΔmtDNA and wild-type mtDNAs (Fig. 3b), however ΔmtDNA was further increased than wild-type mtDNAs, resulting in an increase of ΔmtDNA from 63% to 73%. Consistent with increased ΔmtDNAs, *atfs-1(et18);*ΔmtDNA worms developed slower than *atfs-1(et18)* worms (Fig. 3c) or ΔmtDNA worms[5]. And consistent with further OXPHOS impairment, *atfs-1(et18);*ΔmtDNA worms consumed less oxygen (Fig. 3d), had reduced mitochondrial

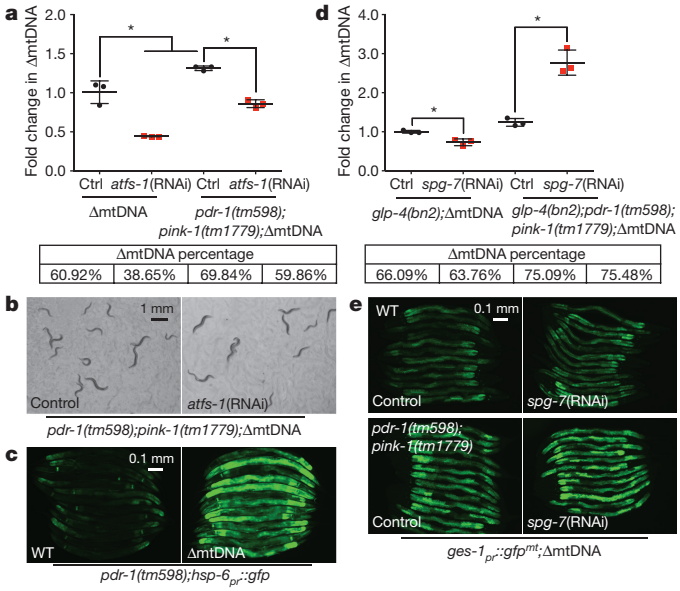


Figure 2 | ATFS-1 promotes ΔmtDNA maintenance and mitochondrial biogenesis largely independent of Parkin. **a**, ΔmtDNA quantification as determined by qPCR in ΔmtDNA and *pdr-1(tm598);pink-1(tm1779);* ΔmtDNA worms raised on control or *atfs-1* RNAi. $n = 3$; error bars, mean ± s.d.; $^*P < 0.03$ (Student's *t*-test). **b**, *pdr-1(tm598);pink-1(tm1779);* ΔmtDNA worms raised on control or *atfs-1* RNAi. Scale bar, 1 mm. **c**, *pdr-1(tm598);hsp-6*$_{pr}$*::gfp* worms with wild-type or ΔmtDNA. Scale bar, 0.1 mm. **d**, ΔmtDNA quantification as determined by qPCR in *glp-4(bn2);*ΔmtDNA and *glp-4(bn2);pdr-1(tm598);pink-1(tm1779);* ΔmtDNA worms raised on control or *spg-7* RNAi. $n = 3$; error bars, mean ± s.d.; $^*P < 0.03$ (Student's *t*-test). **e**, *ges-1*$_{pr}$*::gfp*^{mt};ΔmtDNA or *ges-1*$_{pr}$*::gfp*^{mt}*;pdr-1(tm598);pink-1(tm1779);*ΔmtDNA worms raised on control or *spg-7* RNAi. WT, wild type. Scale bar, 0.1 mm.

© 2016 Macmillan Publishers Limited. All rights reserved

membrane potential (Extended Data Fig. 2c) and were sensitive to the OXPHOS inhibitor rotenone (Extended Data Fig. 2d). Combined, these results indicate that UPR^{mt} activation is detrimental in the presence of ΔmtDNAs.

To investigate potential causes of the OXPHOS deficiency and UPR^{mt} activation in ΔmtDNA worms, we compared expression of a gene within the ΔmtDNA deletion (*ND2*), that is only expressed by the population of wild-type mtDNAs, to mtDNA-encoded genes located outside of the deletion that are encoded by both wild-type and ΔmtDNAs (*ND4*, *ND6*) (Fig. 1a). Surprisingly, *ND2* transcripts were not reduced in ΔmtDNA or *atfs-1(et18);*ΔmtDNA animals despite the absence of the *ND2* gene in ΔmtDNA (Fig. 3e). Interestingly, *ND4* was expressed significantly more in both ΔmtDNA and *atfs-1 (et18);*ΔmtDNA relative to wild-type worms while the nuclear-encoded complex I transcript *nuo-4* was unaffected (Fig. 3e). These results suggest that mitochondrial dysfunction in ΔmtDNA worms is not due to reduction of transcripts encoded by genes within the mtDNA deletion.

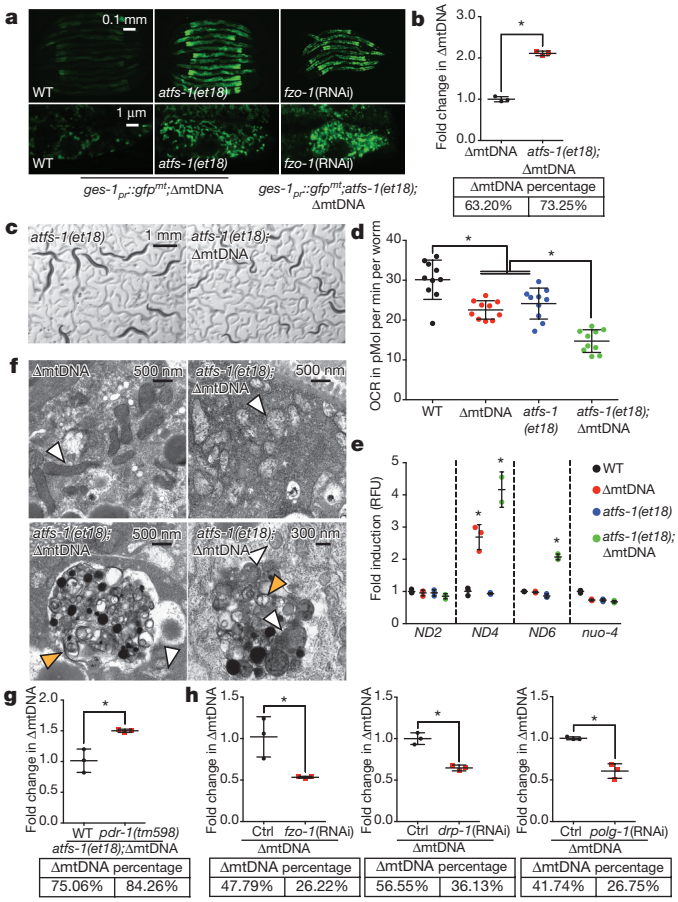


Figure 3 | ATFS-1 activation causes deleterious mtDNA expansion. **a**, *ges-1$_{pr}$::gfp^{mt}*;ΔmtDNA or *ges-1$_{pr}$::gfp^{mt};atfs-1(et18);*ΔmtDNA worms raised on control or *fzo-1* RNAi. The bottom panels are a single intestinal cell. WT, wild type. Scale bar, 0.1 mm (top) and 1 μm (bottom). **b**, ΔmtDNA quantification as determined by qPCR in ΔmtDNA or *atfs-1(et18)*;ΔmtDNA worms. WT, wild type. $n = 3$; error bars, mean ± s.d.; $^*P < 0.03$ (Student's *t*-test). **c**, Synchronized *atfs-1(et18)* and *atfs-1(et18);*ΔmtDNA worms. Scale bar, 1 mm. **d**, Oxygen consumption rates (OCR) of wild-type, ΔmtDNA, *atfs-1(et18)* and *atfs-1(et18)*;ΔmtDNA worms. $n = 10$; error bars, mean ± s.d.; $^*P < 0.03$ (Student's *t*-test). **e**, *ND2*, *ND4*, *ND6 and nuo-4* transcripts as determined by qRT–PCR in wild-type, ΔmtDNA, *atfs-1(et18)* and *atfs-1(et18);*ΔmtDNA worms. $n = 3$; error bars, mean ± s.d.; $^*P < 0.03$ (Student's *t*-test). RFU, relative fluorescent units. **f**, An intestinal cell in ΔmtDNA and *atfs-1(et18);*ΔmtDNA worms. Mitochondria (white) and autophagosomes (orange) are indicated. Scale bars, 500 nm (lower right, 300 nm). **g**, **h**, ΔmtDNA quantification as determined by qPCR in ΔmtDNA, *atfs-1(et18);*ΔmtDNA or *pdr-1(tm598);atfs-1(et18);*ΔmtDNA worms raised on control, *fzo-1*, *drp-1*, or *polg-1* RNAi. $n = 3$; error bars, mean ± s.d.; $^*P < 0.03$ (Student's *t*-test).

Next, we performed electron microscopy to examine mitochondrial morphology in *atfs-1(et18)* harbouring 73% ΔmtDNA. The cristae were largely absent in these mitochondria (Fig. 3f), consistent with severe mitochondrial dysfunction and reduced OXPHOS (Fig. 3d). Notably, a number of autophagosome-like structures associated with degenerate mitochondria were observed only in *atfs-1(et18);*ΔmtDNA worms (Fig. 3f). And, consistent with increased mitophagy in *atfs-1(et18);*ΔmtDNA animals, *pdr-1* deletion further increased the percentage of ΔmtDNA (Fig. 3g). Combined, these results suggest that in the context of deleterious heteroplasmy, increased UPR^{mt} activation can further perturb mitochondrial function, potentially leading to increased mitophagy.

To better understand the ATFS-1-mediated transcriptional program that promotes ΔmtDNA maintenance and propagation, the transcriptomes of wild-type and *atfs-1(et18)* worms were examined. Activated ATFS-1 induced many transcripts suggestive of mitochondrial biogenesis, including the mitochondrial protein import machinery, the cardiolipin synthesis enzyme tafazzin, mitochondrial ribosome and translation factors, prohibitin complex components, mitochondrial chaperones and OXPHOS assembly factors (Supplementary Table 1), consistent with ATFS-1 regulating a mitochondrial biogenesis and mitochondrial proteostasis program to recover mitochondrial function. *atfs-1(et18)* also increased the mtDNA polymerase *polg-1*, the worm orthologue of the mtDNA-binding protein TFAM (*hmg-5*), as well as transcripts required for mitochondrial dynamics (Supplementary Table 1 and Extended Data Fig. 3a).

Interestingly, inhibition of mitochondrial fusion or fission by *fzo-1*(RNAi) or *drp-1*(RNAi) reduced the ΔmtDNA percentage in a similar manner to *atfs-1*(RNAi) (Fig. 3h) in ΔmtDNA worms, suggesting that organelle dynamics stimulated by the UPR^{mt} promote deleterious mtDNA maintenance. Because *fzo-1*(RNAi) inhibition had relatively little effect on mitochondrial biogenesis (Fig. 3a), only organelle morphology (Fig. 3a, lower panel), we speculate that organelle mixing mediated by *drp-1* and *fzo-1*, which requires OXPHOS function[20], limits the enrichment of deleterious mtDNAs in individual organelles and promotes tolerance to deleterious mtDNAs[21,22]. Consistent with this idea, development of *atfs-1(et18);*ΔmtDNA animals was delayed when mitochondrial fusion was inhibited (Extended Data Fig. 3b). Impairment of *polg-1* and *hmg-5* also reduced the ΔmtDNA percentage in ΔmtDNA worms (Fig. 3h and Extended Data Fig. 3c), suggesting that replication and mtDNA protection is involved in maintaining the deleterious genome. Combined, these data suggest that through multiple outputs, ATFS-1 activation provides favourable conditions for ΔmtDNA proliferation.

These results suggest an unanticipated consequence of ATFS-1 and UPR^{mt} activation in the context of deleterious mtDNA heteroplasmy. While the UPR^{mt} is protective during exposure to mitochondrial toxins[23] or mutations within nuclear-encoded OXPHOS genes[4,24], in the context of mtDNA heteroplasmy, UPR^{mt} activation maintains the deleterious mtDNA. We propose that by inducing a mitochondrial recovery program, UPR^{mt} activation inadvertently propagates deleterious mtDNAs in an attempt to recover OXPHOS activity. These results potentially shed light on the underlying mechanisms that lead to mitochondrial diseases and the enrichment of ΔmtDNAs found in aged cells[7,8]. They also emphasize the importance of UPR^{mt} regulation and suggest that prolonged UPR^{mt} activation is potentially harmful[25], as ATFS-1 activation creates an environment favourable for ΔmtDNAs.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 25 November; accepted 8 April 2016.
Published online 2 May 2016.

© 2016 Macmillan Publishers Limited. All rights reserved

1. Bua, E. *et al.* Mitochondrial DNA-deletion mutations accumulate intracellularly to detrimental levels in aged human skeletal muscle fibers. *Am. J. Hum. Genet.* **79,** 469–480 (2006).
2. Kraytsberg, Y. *et al.* Mitochondrial DNA deletions are abundant and cause functional impairment in aged human substantia nigra neurons. *Nat. Genet.* **38,** 518–520 (2006).
3. Nargund, A. M., Fiorese, C. J., Pellegrino, M. W., Deng, P. & Haynes, C. M. Mitochondrial and nuclear accumulation of the transcription factor ATFS-1 promotes OXPHOS recovery during the UPR(mt). *Mol. Cell* **58,** 123–133 (2015).
4. Nargund, A. M., Pellegrino, M. W., Fiorese, C. J., Baker, B. M. & Haynes, C. M. Mitochondrial import efficiency of ATFS-1 regulates mitochondrial UPR activation. *Science* **337,** 587–590 (2012).
5. Tsang, W. Y. & Lemire, B. D. Stable heteroplasmy but differential inheritance of a large mitochondrial DNA deletion in nematodes. *Biochem. Cell Biol.* **80,** 645–654 (2002).
6. Greaves, L. C. *et al.* Clonal expansion of early to mid-life mitochondrial DNA point mutations drives mitochondrial dysfunction during human ageing. *PLoS Genet.* **10,** e1004620 (2014).
7. Taylor, S. D. *et al.* Targeted enrichment and high-resolution digital profiling of mitochondrial DNA deletions in human brain. *Aging Cell* **13,** 29–38 (2014).
8. Stewart, J. B. & Chinnery, P. F. The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. *Nat. Rev. Genet.* **16,** 530–542 (2015).
9. Yoneda, T. *et al.* Compartment-specific perturbation of protein handling activates genes encoding mitochondrial chaperones. *J. Cell Sci.* **117,** 4055–4066 (2004).
10. Durieux, J., Wolff, S. & Dillin, A. The cell-non-autonomous nature of electron transport chain-mediated longevity. *Cell* **144,** 79–91 (2011).
11. Baker, B. M., Nargund, A. M., Sun, T. & Haynes, C. M. Protective coupling of mitochondrial function and protein synthesis via the eIF2α kinase GCN-2. *PLoS Genet.* **8,** e1002760 (2012).
12. Haynes, C. M., Petrova, K., Benedetti, C., Yang, Y. & Ron, D. ClpP mediates activation of a mitochondrial unfolded protein response in *C. elegans*. *Dev. Cell* **13,** 467–480 (2007).
13. Narendra, D. P. *et al.* PINK1 is selectively stabilized on impaired mitochondria to activate Parkin. *PLoS Biol.* **8,** e1000298 (2010).
14. Valenci, I., Yonai, L., Bar-Yaacov, D., Mishmar, D. & Ben-Zvi, A. Parkin modulates heteroplasmy of truncated mtDNA in *Caenorhabditis elegans*. *Mitochondrion* **20,** 64–70 (2015).
15. Suen, D. F., Narendra, D. P., Tanaka, A., Manfredi, G. & Youle, R. J. Parkin overexpression selects against a deleterious mtDNA mutation in heteroplasmic cybrid cells. *Proc. Natl Acad. Sci. USA* **107,** 11835–11840 (2010).
16. Scarffe, L. A., Stevens, D. A., Dawson, V. L. & Dawson, T. M. Parkin and PINK1: much more than mitophagy. *Trends Neurosci.* **37,** 315–324 (2014).
17. Jin, S. M. & Youle, R. J. The accumulation of misfolded proteins in the mitochondrial matrix is sensed by PINK1 to induce PARK2/Parkin-mediated mitophagy of polarized mitochondria. *Autophagy* **9,** 1750–1757 (2013).
18. Palikaras, K., Lionaki, E. & Tavernarakis, N. Coordination of mitophagy and mitochondrial biogenesis during ageing in *C. elegans*. *Nature* **521,** 525–528 (2015).
19. Rauthan, M., Ranji, P., Aguilera Pradenas, N., Pitot, C. & Pilon, M. The mitochondrial unfolded protein response activator ATFS-1 protects cells from inhibition of the mevalonate pathway. *Proc. Natl Acad. Sci. USA* **110,** 5981–5986 (2013).
20. Mishra, P., Carelli, V., Manfredi, G. & Chan, D. C. Proteolytic cleavage of Opa1 stimulates mitochondrial inner membrane fusion and couples fusion to oxidative phosphorylation. *Cell Metab.* **19,** 630–641 (2014).
21. Chen, H. *et al.* Mitochondrial fusion is required for mtDNA stability in skeletal muscle and tolerance of mtDNA mutations. *Cell* **141,** 280–289 (2010).
22. Nakada, K. *et al.* Inter-mitochondrial complementation: Mitochondria-specific system preventing mice from expression of disease phenotypes by mutant mtDNA. *Nat. Med.* **7,** 934–940 (2001).
23. Pellegrino, M. W. *et al.* Mitochondrial UPR-regulated innate immunity provides resistance to pathogen infection. *Nature* **516,** 414–417 (2014).
24. Schieber, M. & Chandel, N. S. TOR signaling couples oxygen sensing to lifespan in *C. elegans*. *Cell Reports* **9,** 9–15 (2014).
25. Lamech, L. T. & Haynes, C. M. The unpredictability of prolonged activation of stress response pathways. *J. Cell Biol.* **209,** 781–787 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank B. Lemire, the National Bioresource Project and the *Caenorhabditis* Genetics Center for providing *C. elegans* strains (funded by NIH Office of Research 362 Infrastructure Programs (P40 OD010440), and the Genomics and Bioinformatics Facilities at Memorial Sloan Kettering Cancer Center). This work was supported by National Institutes of Health grants (R01AG040061 and R01AG047182) to C.M.H. and (R01HD078703 and R01NS081490) to S.S., and a Parkinson's Disease Foundation grant (PDF-FBS-1314) to Y.-F.L. and Deutsche Forschungsgemeinschaft (DFG, SCHU 3023/1-1) to A.M.S.

Author Contributions Y.-F.L. and C.M.H. planned the experiments. Y.-F.L. generated the worm strains and performed the mtDNA quantification and obtained the images. A.M.S. and Y.-F.L. performed the oxygen consumption analysis and M.W.P. performed the microarray experiments. Electron microscopy was performed by Y.Lu under the supervision of S.S. Y.-F.L. and C.M.H. wrote the manuscript.

Author Information The microarray data have been deposited in a MIAME-compliant format to the Gene Expression Omnibus database under accession number GSE73669. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to C.M.H. (haynesc@mskcc.org).

© 2016 Macmillan Publishers Limited. All rights reserved

METHODS

Worm strains, staining and rotenone treatment. The *atfs-1(tm4919)*, *pink-1(tm1779)* and *pdr-1(tm598)* strains were obtained from the National BioResource Project (Tokyo, Japan) and the N2 (wild-type mtDNA), LB138 (*uaDf5* or ΔmtDNA), *atg-18(gk378)*, *isp-1(qm150)* and *glp-4(bn2)* strains from the *Caenorhabditis* Genetics Center (Minneapolis, Minnesota). The *atfs-1(et18)* strain was a gift from M. Pilon. The *ges-1*$_{pr}$*::gfp*^{mt} worms used for visualizing mitochondrial content and the reporter strain *hsp-6*$_{pr}$*::gfp* for visualizing UPR^{mt} activation have been described[9,23]. Hermaphrodites were used for all experiments. Rotenone treatment and TMRE staining was performed by synchronizing and raising worms on plates previously soaked with M9 buffer containing rotenone (Tocris), DMSO, or TMRE (Molecular Probes). TMRE-stained worms were subsequently placed on control plates to remove TMRE-containing bacteria in the digestive tract for 3 h before imaging.

mtDNA quantification. Wild-type mtDNA and ΔmtDNA quantification was performed using qPCR-based methods similar to previously described assays[14,26]. 20–30 worms were collected in 30 μl of lysis buffer (50 mM KCl, 10 mM Tris-HCl (pH 8.3), 2.5 mM $MgCl_2$, 0.45% NP-40, 0.45% Tween 20, 0.01% gelatin, with freshly added 200 μg ml^{-1} proteinase K) and frozen at −80 °C for 20 min before lysis at 65 °C for 80 min. Relative quantification was used for determining the fold changes in mtDNA between samples. 1 μl of lysate was used in each triplicate qPCR reaction. qPCR was performed using the Thermo-Scientific SyBr Green Maxima Mix and the MyiQ2 Two-Colour Real-Time PCR Detection System (Bio-Rad Laboratories). Primers that specifically amplify wild-type or ΔmtDNA are listed in Supplementary Table 2. Primers that amplify a non-coding region near the nuclear-encoded *ges-1* gene were used as a control (Supplementary Table 2).

Absolute quantification was also performed to determine the percentage or ratio of ΔmtDNA relative to total mtDNA (ΔmtDNA and wild-type mtDNA). This approach was also employed to determine ΔmtDNA number per nuclear genome in the individual worm analysis. Standard curves for each qPCR primer set were generated using purified plasmids individually containing approximately 1 kb of the mtDNA or nuclear DNA sequence specific for each primer set[14,26]. The concentration of the purified plasmids was determined using a Nanodrop spectrometer (Thermo Scientific). mtDNA was harvested from synchronized worms at the L4 stage. For the experiments involving the *glp-4(bn2)* strain, worms were raised at 25 °C and harvested 3 days after hatching. All qPCR results are presented as technical replicates, but each experiment has been repeated three or more times. A Student's *t*-test was employed to determine the level of statistical significance.

RNA isolation and qRT–PCR analysis. Worms were synchronized and raised in liquid culture until the L4 stage when they were harvested and compacted into pellets on ice. Total RNA was extracted from a 30–50 μl worm pellet using RNA STAT (Tel-Test). For the analysis of mtDNA-encoded mRNAs, the RNA extracts were then treated with DNase using the DNA-free kit (Ambion) to reduce mtDNA contamination. 1 μg of RNA was used for synthesizing cDNA with the iScript cDNA Synthesis Kit (Bio-Rad Laboratories). qPCR was performed using the Thermo-Scientific SyBr Green Maxima Mix and the MyiQ2 Two-Colour Real-Time PCR Detection System (Bio-Rad Laboratories). Primer sequences are listed in Supplementary Table 2. All qPCR results are presented as technical replicates, but each experiment has been repeated three or more times. A Student's *t*-test was employed to determine the level of statistical significance.

Statistics. All experiments were performed three times yielding similar results and comprised of biological replicates. The sample size and statistical tests were chosen based on previous studies with similar methodologies and the data met the assumptions for each statistical test performed. No statistical method was used in deciding sample sizes. No blinded experiments were performed and randomization was not used. For all figures, the mean ± standard deviation (s.d.) is represented unless otherwise noted.

Imaging. Whole-worm images were obtained using either a Zeiss AxioCam MRm camera mounted on a Zeiss Imager Z2 microscope or a Zeiss M2BIO dissecting scope. Exposure times were the same within each experiment. Fluorescent (*ges-1*$_{pr}$*::gfp*^{mt}) mitochondrial morphology images of the most proximal intestinal cells were taken with a Nikon Eclipse Ti confocal microscope. The signal intensity of each image was adjusted to highlight the differences in mitochondrial morphology as opposed to mitochondrial density. All images are representative of more than three images.

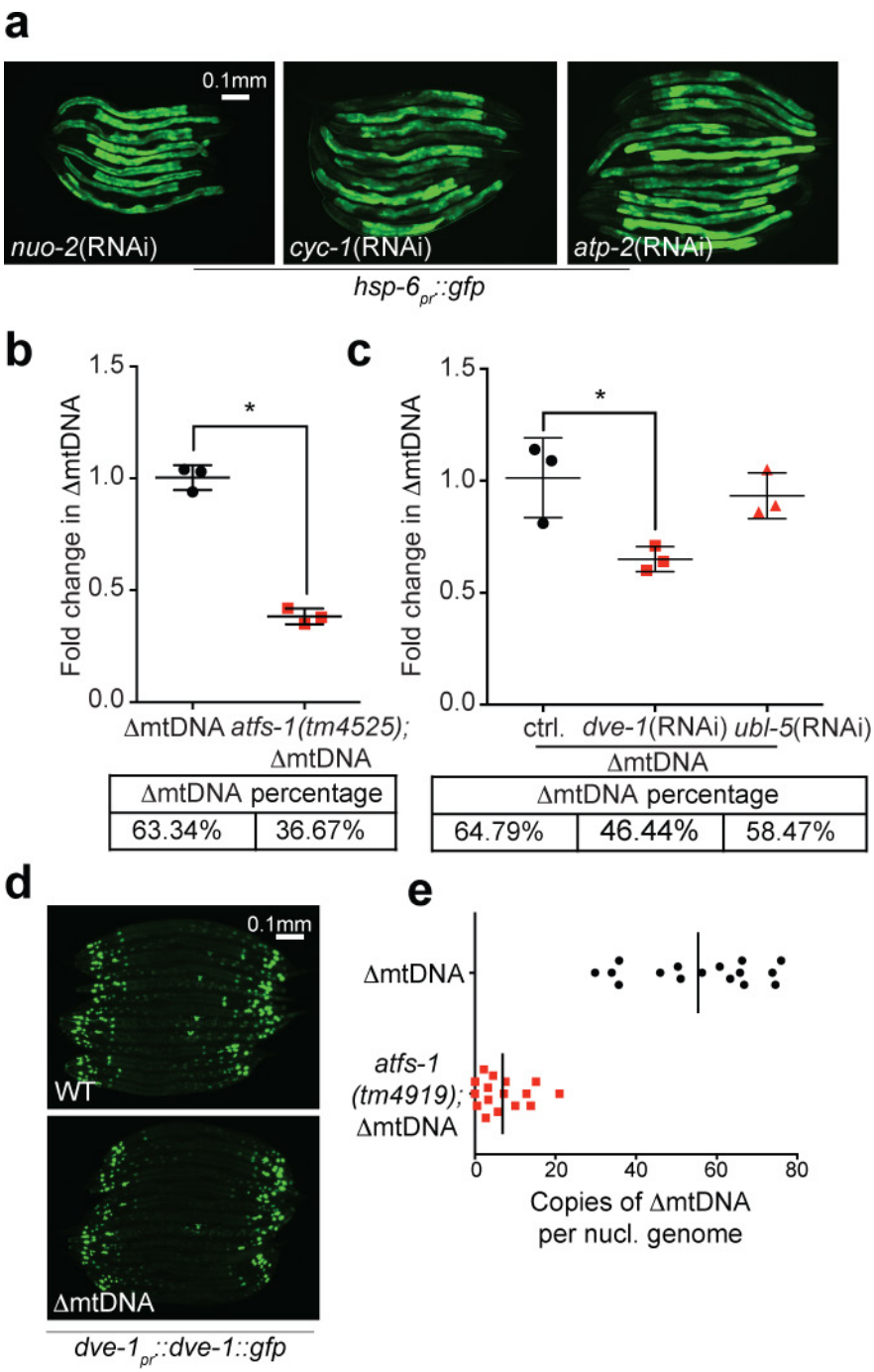
Electron microscopy. Animals were prepared for electron microscopy using standard methods[27]. Ultrathin serial sections (80 nm) were collected using a Leica Ultracut UCT Ultramicrotome. Sections at two levels, 100 μm and 110 μm away from the head region, for each genotype were examined. Electron microscopy images were acquired using a FEI Tecnai G2 Spirit BioTwin transmission electron microscope operating at 120 kV with a Gatan 4 K × 4 K digital camera.

Microarray analysis. Wild-type (N2) and *atfs-1(et18)* worms were synchronized by bleaching, and harvested at the L4 stage of development. Total RNA was extracted using the RNA STAT reagent (Tel-Test) and used for double-stranded cDNA synthesis using the iScript cDNA Synthesis Kit (Bio-Rad Laboratories). Microarray analysis using GeneChip *C. elegans* genome arrays (Affymetrix) was conducted as previously described[4]. Differences in gene expression between wild-type and *atfs-1(et18)* worms was determined using Anova streamlined (Partek Genomic Suite (v6.5)). Supplementary Table 1 contains the relative fold induction and *P* values for each mRNA. Confirmation of the microarray results was performed via qRT–PCR as described[4]. Primers are listed in Supplementary Table 2.

Oxygen consumption. Oxygen consumption was measured using a Seahorse XF^{e}96 Analyzer at 25 °C similar to that described previously[28]. In brief, adult worms were transferred onto empty plates and allowed to completely digest the remaining bacteria for 1 h, after which four worms were transferred into each well of a 96-well microplate containing 200 μl M9 buffer. Basal respiration was measured for a total of 60 min, in 6-min intervals that included a 2 min mix, a 2 min time delay and a 2 min measurement. To measure respiratory capacity, 15 μM FCCP was injected, the OCR (oxygen consumption rate) reading was allowed to stabilize for 15 min then measured for five consecutive intervals. Each measurement was considered one technical replicate.

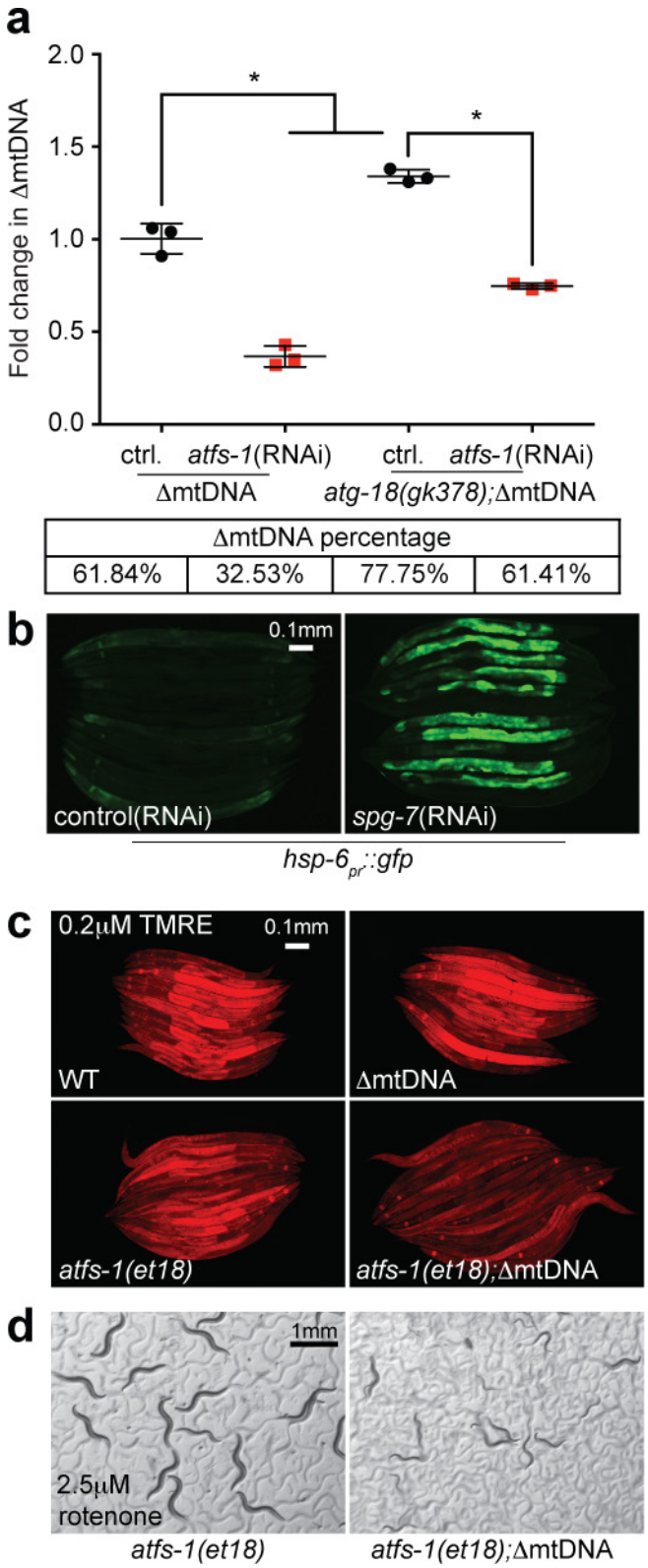
26. He, L. *et al.* Detection and quantification of mitochondrial DNA deletions in individual cells by real-time PCR. *Nucleic Acids Res.* **30,** e68 (2002).
27. Bargmann, C. I., Hartwieg, E. & Horvitz, H. R. Odorant-selective genes and neurons mediate olfaction in *C. elegans*. *Cell* **74,** 515–527 (1993).
28. Andreux, P. A. *et al.* A method to identify and validate mitochondrial modulators using mammalian cells and the worm *C. elegans*. *Sci. Rep.* **4,** 5285 (2014).

© 2016 Macmillan Publishers Limited. All rights reserved



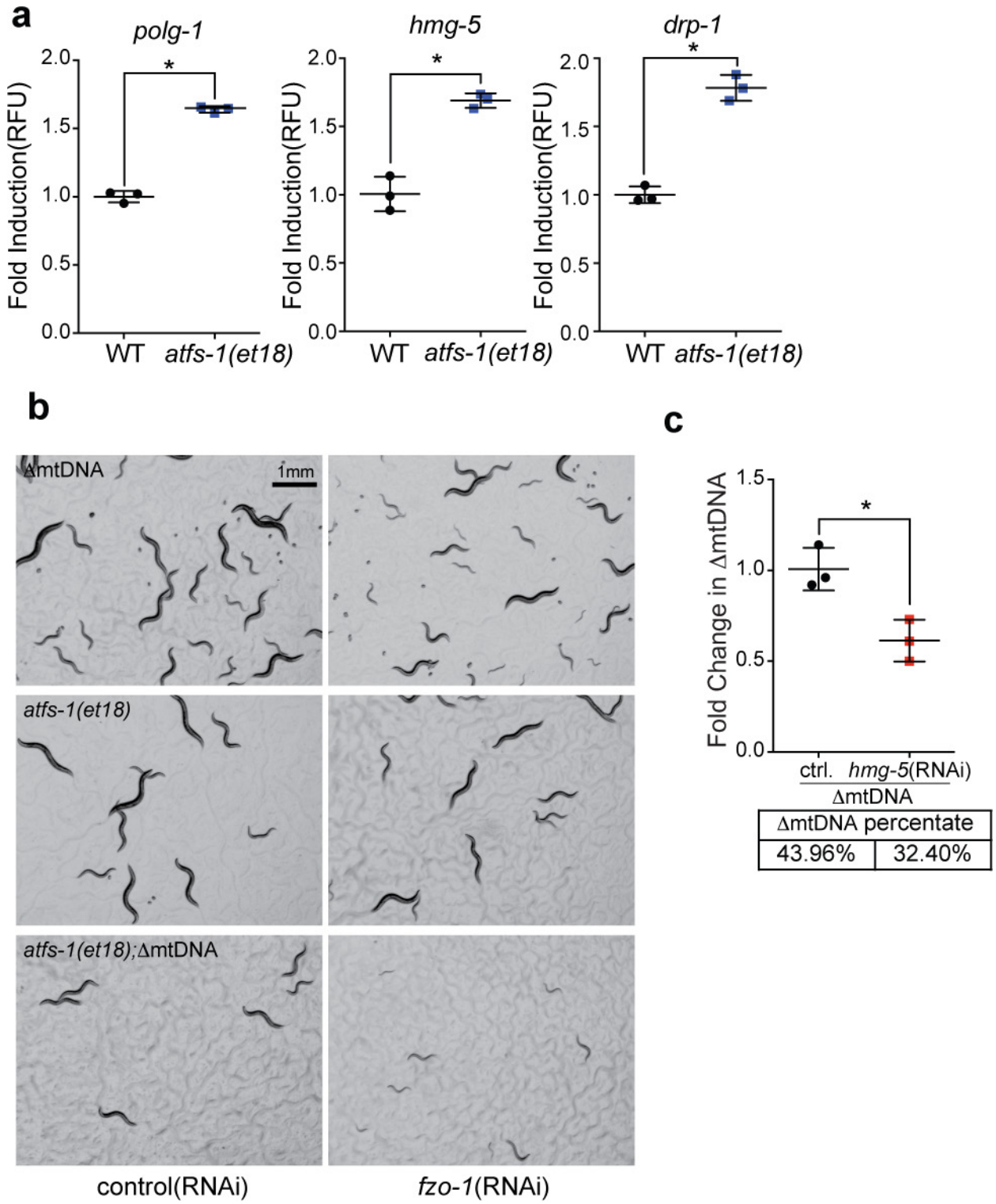
Extended Data Figure 1 | RNAi of nuclear encoded OXPHOS components activates the UPR^{mt}, and UPR^{mt} signalling components are required for ΔmtDNA propagation. a, Images of *hsp-6$_{pr}$::gfp* worms raised on *nuo-2* (complex I), *cyc-1* (cytochrome *c*) or *atp-2* (complex V) RNAi. Scale bar, 0.1 mm. **b**, Relative ΔmtDNA quantification as determined by qPCR in ΔmtDNA or *atfs-1(tm4525);*ΔmtDNA worms. $n = 3$; error bars, mean ± s.d.; $^*P < 0.03$ (Student's *t*-test). **c**, Relative ΔmtDNA quantification as determined by qPCR in ΔmtDNA worms raised on control, *dve-1*, or *ubl-5* RNAi. $n = 3$; error bars, mean ± s.d.; $^*P < 0.03$ (Student's *t*-test). **d**, Images of *dve-1$_{pr}$::dve-1::gfp* in wild-type or ΔmtDNA worms. Scale bar, 0.1 mm. **e**, Ratios of ΔmtDNA and nuclear genomic DNA as determined by qPCR in individual ΔmtDNA and *atfs-1(tm4919);*ΔmtDNA worms. Black bars represent the mean. $n = 16$; mean $P < 0.0001$ (Student's *t*-test).

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 2 | The reduction in ΔmtDNA caused by *atfs-1*(RNAi) is largely independent of autophagy, and ATFS-1 activation in the presence of a deleterious mtDNA is harmful. **a**, Relative ΔmtDNA quantification as determined by qPCR in ΔmtDNA and *atg-18(gk378);*ΔmtDNA worms raised on control or *atfs-1* RNAi. $n = 3$; error bars, mean ± s.d.; $^*P < 0.03$ (Student's *t*-test). **b**, Images of *hsp-6*$_{pr}$*::gfp* worms raised on control or *spg-7* RNAi. Scale bar, 0.1 mm. **c**, Images of TMRE-stained wild-type, ΔmtDNA, *atfs-1(et18)* and *atfs-1 (et18);*ΔmtDNA worms. Scale bar, 0.1 mm. **d**, Images of synchronized *atfs-1(et18)* and *atfs-1(et18);*ΔmtDNA worms raised on 2.5 μM rotenone 4 days after hatching. Scale bar, 1 mm.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 3 | *polg-1*, *hmg-5* and *drp-1* mRNAs are induced in *atfs-1(et18)* worms, and mitochondrial fusion is required for the development of *atfs-1(et18)* worms harbouring ΔmtDNA. a, *polg-1*, *hmg-5* and *drp-1* transcripts as determined by qRT–PCR in wild-type and *atfs-1(et18)* worms. $n = 3$; error bars, mean ± s.d.; $^*P < 0.03$ (Student's *t*-test). RFU, random fluorescent units. **b**, Images of synchronized ΔmtDNA, *atfs-1(et18)*, and *atfs-1(et18)*;ΔmtDNA worms raised on control or *fzo-1* RNAi. Scale bar, 1 mm. **c**, ΔmtDNA quantification as determined by qPCR in ΔmtDNA worms raised on control or *hmg-5* RNAi. $n = 3$; error bars, mean ± s.d.; $^*P < 0.03$ (Student's *t*-test).

© 2016 Macmillan Publishers Limited. All rights reserved

LETTER

doi:10.1038/nature17946

Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage

Alexis C. Komor[1,2], Yongjoo B. Kim[1,2], Michael S. Packer[1,2], John A. Zuris[1,2] & David R. Liu[1,2]

Current genome-editing technologies introduce double-stranded (ds) DNA breaks at a target locus as the first step to gene correction[1,2]. Although most genetic diseases arise from point mutations, current approaches to point mutation correction are inefficient and typically induce an abundance of random insertions and deletions (indels) at the target locus resulting from the cellular response to dsDNA breaks[1,2]. Here we report the development of 'base editing', a new approach to genome editing that enables the direct, irreversible conversion of one target DNA base into another in a programmable manner, without requiring dsDNA backbone cleavage or a donor template. We engineered fusions of CRISPR/Cas9 and a cytidine deaminase enzyme that retain the ability to be programmed with a guide RNA, do not induce dsDNA breaks, and mediate the direct conversion of cytidine to uridine, thereby effecting a C→T (or G→A) substitution. The resulting 'base editors' convert cytidines within a window of approximately five nucleotides, and can efficiently correct a variety of point mutations relevant to human disease. In four transformed human and murine cell lines, second- and third-generation base editors that fuse uracil glycosylase inhibitor, and that use a Cas9 nickase targeting the non-edited strand, manipulate the cellular DNA repair response to favour desired base-editing outcomes, resulting in permanent correction of ∼15–75% of total cellular DNA with minimal (typically ≤1%) indel formation. Base editing expands the scope and efficiency of genome editing of point mutations.

The clustered regularly interspaced short palindromic repeat (CRISPR) system has been widely used to mediate genome editing in a variety of organisms and cell lines[3]. CRISPR/Cas9 protein–RNA complexes localize to a target DNA sequence through base pairing with a guide RNA, and natively create a dsDNA break (DSB) at the locus specified by the guide RNA. In response to DSBs, cellular DNA repair processes mostly result in random insertions or deletions (indels) at the site of DNA cleavage through non-homologous end joining (NHEJ). In the presence of a homologous DNA template, the DNA surrounding the cleavage site can be replaced through homology-directed repair (HDR). HDR competes with NHEJ during the resolution of DSBs, and indels are generally more abundant outcomes than gene replacement. For most known genetic diseases, however, correction of a point mutation in the target locus, rather than stochastic disruption of the gene, is needed to study or address the underlying cause of the disease[4].

Motivated by this need, researchers have sought to increase the efficiency of HDR and suppress NHEJ. Despite recent progress (see Supplementary Information for a detailed discussion), current strategies to correct point mutations using HDR under therapeutically relevant conditions remain inefficient (typically ∼0.1 to 5%)[5,6], especially in unmodified, non-dividing cells. These observations highlight the need to develop alternative approaches to correct point mutations in genomic DNA that do not require DSBs.

We envisioned that direct conversion of one DNA base to another at a programmable target locus without requiring DSBs could increase the efficiency of gene correction relative to HDR without introducing an excess of random indels. Catalytically dead Cas9 (dCas9), which contains Asp10Ala and His840Ala mutations that inactivate its nuclease activity, retains its ability to bind DNA in a guide RNA-programmed manner, but does not cleave the DNA backbone[7]. In principle, conjugation of dCas9 with an enzymatic or chemical catalyst that mediates the direct conversion of one base to another could enable RNA-programmed DNA base editing.

The deamination of cytosine (C) is catalysed by cytidine deaminases[8] and results in uracil (U), which has the base-pairing properties of thymine (T). Most known cytidine deaminases operate on RNA, and the few examples that are known to accept DNA require single-stranded (ss) DNA[9]. Recent studies on the dCas9–target DNA complex reveal that at least nine nucleotides (nt) of the displaced DNA strand are unpaired upon formation of the Cas9–guide RNA–DNA 'R-loop' complex[10]. Indeed, in the structure of the Cas9 R-loop complex, the first 11 nt of the protospacer on the displaced DNA strand are disordered, suggesting that their movement is not highly restricted[11]. It has also been speculated that Cas9 nickase-induced mutations at cytosines in the non-template strand might arise from their accessibility by cellular cytosine deaminase enzymes[12]. We reasoned that a subset of this stretch of ssDNA in the R-loop might serve as an efficient substrate for a dCas9-tethered cytidine deaminase to effect direct, programmable conversion of C to U in DNA (Fig. 1a).

Four different cytidine deaminase enzymes (human AID, human APOBEC3G, rat APOBEC1, and lamprey CDA1) were evaluated for ssDNA deamination. Of the four enzymes, rat APOBEC1 showed the highest deaminase activity under the conditions tested (Extended Data Fig. 1a). Fusing rat APOBEC1 to the amino terminus, but not the carboxy terminus, of dCas9 preserves deaminase activity (Extended Data Fig. 1a). We expressed and purified four rat APOBEC1–dCas9 fusions with linkers of different length and composition (Extended Data Fig. 1b), and evaluated each fusion for single guide RNA (sgRNA)-programmed dsDNA deamination *in vitro* (Fig. 1b and Extended Data Fig. 1c–f).

We observed efficient, sequence-specific, sgRNA-dependent C to U conversion *in vitro* (Fig. 1c). Conversion efficiency was greatest using rat APOBEC1–dCas9 linkers over nine amino acids in length. The number of positions susceptible to deamination (the 'activity window') increases from approximately 3 to 6 nt as the linker length was extended from 3 to 21 amino acids (Extended Data Fig. 1c–f). The 16-residue XTEN linker[13] offered a promising balance between these two characteristics, with an efficient deamination window of approximately 5 nt, typically from positions 4 to 8 within the protospacer, counting the end distal to the protospacer-adjacent motif (PAM) as position 1. The rat APOBEC1–XTEN–dCas9 protein served as the first-generation base editor (BE1).

We assessed the ability of BE1 *in vitro* to correct seven T→C mutations relevant to human disease (Extended Data Fig. 2). BE1 yielded products consistent with efficient editing of the target C, or of at least

[1]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [2]Howard Hughes Medical Institute, Harvard University, Cambridge, Massachusetts 02138, USA.

© 2016 Macmillan Publishers Limited. All rights reserved

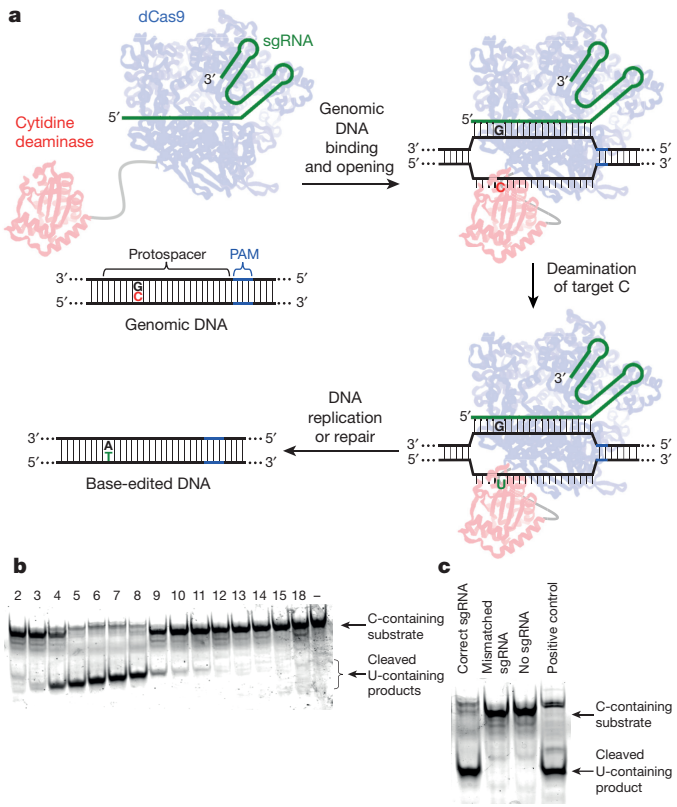


Figure 1 | First-generation base editor (BE1) mediates specific, guide RNA-programmed C→U conversion *in vitro*. **a**, Base editing strategy. DNA with a target C (red) at a locus specified by a guide RNA (green) is bound by dCas9 (blue), which mediates local DNA strand separation. Cytidine deamination by a tethered APOBEC1 enzyme (red) converts the single-stranded target C→U. The resulting G:U heteroduplex can be permanently converted to an A:T base pair following DNA replication or DNA repair. **b**, Deamination assay showing a BE1 activity window of approximately five nucleotides. Samples were prepared as described in the Methods. Each lane is labelled according to the position of the target C within the protospace, or with '–' if no target C is present, counting the base distal from the PAM as position 1. **c**, Deamination assay showing the sequence specificity and sgRNA-dependence of BE1. The DNA substrate with C at position 7 in **b** was incubated with BE1 and the correct sgRNA, a mismatched sgRNA or no sgRNA. The positive control sample used a synthetic DNA substrate with a U at position 7. For uncropped gel data, see Supplementary Figure 1.

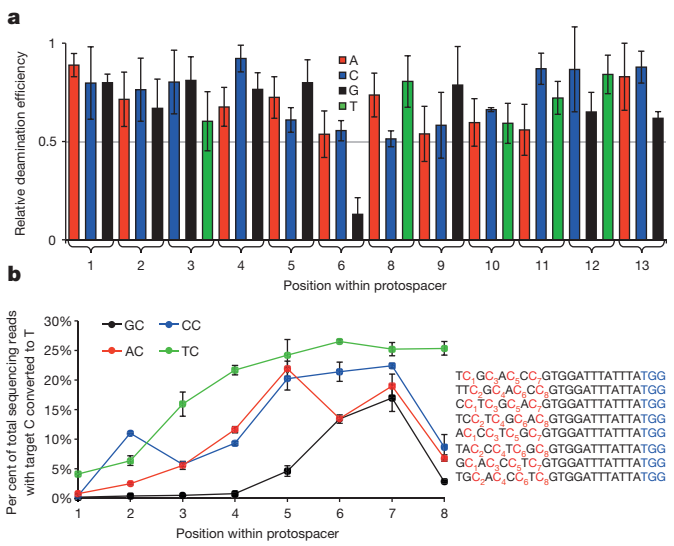


Figure 2 | Effects of sequence context and target C position on base editing efficiency *in vitro*. **a**, Effect of changing the sequence surrounding the target C on editing efficiency *in vitro*. The deamination yield of 80% of targeted strands (40% of total sequencing reads from both strands) for C_7 in the protospacer sequence 5′-TTATTT(C_7)GTGGATTTATTTA-3′ was defined as 1.0, and the relative deamination efficiencies of substrates containing all possible single-base mutations at positions 1–6 and 8–13 are shown. **b**, Positional effect of each NC motif on editing efficiency *in vitro*. Each NC target motif was varied from positions 1 to 8 within the protospacer as indicated in the sequences shown on the right. The PAM is shown in blue. The graph shows the percentage of total DNA sequencing reads containing T at each of the numbered target C positions following incubation with BE1. Note that the maximum possible deamination yield *in vitro* is 50% of total sequencing reads (100% of targeted strands). Values and error bars reflect the mean and standard deviation of three (**a**) or two (**b**) independent biological replicates performed on different days.

one C within the activity window when multiple Cs were present, in six of these seven targets *in vitro*, with an average apparent editing efficiency of 44% (Extended Data Fig. 2).

Although the preferred sequence context for APOBEC1 substrates is TC or CC[14], we anticipated that the increased effective molarity of the tethered deaminase and its ssDNA substrate upon dCas9 binding might relax this preference. To illuminate the context dependence of BE1, we assayed its ability to edit a dsDNA 60-mer containing a single fixed C at position 7 within the protospacer, as well as all 36 single-mutant variants in which protospacer bases 1–6 and 8–13 were individually varied to each of the other three bases. High-throughput DNA sequencing revealed 50–80% C to U conversion of substrate strands (25–40% of sequence reads from both DNA strands, one of which is not a substrate for BE1) (Fig. 2a). Editing efficiency was independent of sequence context, unless the base immediately 5′ of the target C was a G, in which case editing efficiency was substantially lower (Fig. 2a). Next we assessed BE1 activity *in vitro* on all four NC motifs at positions 1 through 8 within the protospacer (Fig. 2b). BE1 activity followed the order TC ≥ CC ≥ AC > GC (the second nucleotide (C) is the target nucleotide), with maximum editing efficiency achieved when the target C is at or near position 7 (see Supplementary Information). In addition, we observed that the base editor is processive, and will efficiently convert most or all Cs to Us on the same DNA strand within the five-base activity window (Extended Data Fig. 3).

While BE1 efficiently processes substrates in a test tube, in cells, a variety of possible DNA repair outcomes determines the fate of the initial U:G product of base editing (Fig. 3a). We tested the ability of BE1 to convert C→T in human cells on 14 Cs in six well-studied target sites in the human genome (see Supplementary Information and Extended Data Fig. 4a)[15]. Although C→T editing in cells was observed for all cases, the efficiency of base editing was 0.8% to 7.7% of total DNA sequences, a large 5- to 36-fold decrease in efficiency compared to that of *in vitro* base editing (Fig. 3b and Extended Data Fig. 4).

We hypothesized that the cellular DNA repair response to U:G heteroduplex DNA was responsible for the large decrease in base editing efficiency in cells (Fig. 3a). Uracil DNA glycosylase (UDG) catalyses removal of U from DNA in cells and initiates base-excision repair (BER), with reversion of the U:G pair to a C:G pair as the most common outcome (Fig. 3a)[16]. Uracil DNA glycosylase inhibitor (UGI), an 83-residue protein from *Bacillus subtilis* bacteriophage PBS1, potently blocks human UDG activity (IC_{50} = 12 pM)[17]. In an effort to subvert BER at the site of base editing, we fused UGI to the C terminus of BE1 to create a second-generation base editor (BE2, APOBEC–XTEN–dCas9–UGI) and repeated editing assays on all six genomic loci. Editing efficiencies in human cells were on average threefold higher with BE2 than BE1, resulting in gene conversion efficiencies of up to 20% of total DNA sequenced (Fig. 3b).

Importantly, BE1 and BE2 resulted in indel formation rates ≤ 0.1% (Fig. 3c and Extended Data Table 1), consistent with the known mechanistic dependence of NHEJ on DSBs (see Supplementary Information)[18]. We assessed BE2-mediated base editing efficiencies on the same genomic targets in U2OS cells, and observed results similar

© 2016 Macmillan Publishers Limited. All rights reserved

to those in HEK293T cells (Extended Data Fig. 5). Together, these results indicate that conjugating UGI to BE1 can increase the efficiency of base editing in human cells.

Converting and protecting the substrate strand of a C:G base pair (bp) results in a maximum base editing yield of 50%. To augment base editing efficiency beyond this limit, we sought to further manipulate cellular DNA repair to induce correction of the non-edited strand containing the G. Eukaryotic mismatch repair (MMR) uses nicks present in newly synthesized DNA to direct removal and resynthesis of the newly synthesized strand (Fig. 3a)[19,20]. We reasoned that nicking the DNA strand containing the unedited G would simulate newly synthesized DNA, inducing MMR, or simulate damaged DNA, inducing long-patch BER, to preferentially resolve the U:G mismatch into desired U:A and T:A products (Fig. 3a). We therefore restored the catalytic His residue at position 840 in the Cas9 HNH domain of BE2 (ref. 7), resulting in the third-generation base editor (BE3, APOBEC–XTEN–dCas9(A840H)–UGI) that nicks the non-edited strand containing a G opposite the edited U. BE3 retains the Asp10Ala mutation in Cas9 that prevents dsDNA cleavage, and also retains UGI to suppress UDG-initiated BER of the editing strand.

Nicking the non-edited strand augmented base editing efficiency in human cells treated with BE3 by an additional two- to sixfold relative to BE2, resulting in up to 37% of total DNA sequences containing the targeted C→T conversion (Fig. 3b). Importantly, only a small frequency of indels, averaging 1.1% for the six tested loci, was observed from BE3 treatment (Fig. 3c and Extended Data Table 1). In contrast, when we treated cells with wild-type Cas9, sgRNA to target each of three loci, and a 200 nt ssDNA donor template to mediate HDR, we observed C→T conversion efficiencies averaging only 0.5%, with much higher indel formation averaging 4.3% (Fig. 3c). The ratio of allele conversion to NHEJ outcomes averaged >1,000 for BE2, 23 for BE3, and 0.17 for wild-type Cas9 (Fig. 3c). We confirmed the permanence of base editing in human cells by monitoring editing efficiencies over multiple cell divisions in HEK293T cells at the HEK293 site 3 and 4 genomic loci (Extended Data Fig. 6 and Supplementary Information). These results collectively establish that base editing can induce much more efficient targeted single-base editing in human cells than Cas9-mediated HDR, and with substantially less (BE3) or almost no (BE2) indel formation.

Next, we examined the off-target activity of BE1, BE2, and BE3 in human cells for five previously studied loci (see Supplementary Information and Supplementary Tables 1–5). Because the sequence preference of rat APOBEC1 is known to be independent of bases more than one nucleotide from the target C (ref. 21), consistent with Fig. 2a, we assumed that off-target base editing arises from off-target Cas9 binding. Therefore, we sequenced the top 34 known Cas9 off-target sites in human cells[15], and the top 12 known dCas9 off-target binding sites for the six genomic loci studied in Fig. 3 (Supplementary Tables 1–5)[22]. We observed detectable off-target base editing at a subset of known Cas9 off-target sites (16/34 for BE1 and BE2; and 17/34 for BE3), but no detectable base editing at the known dCas9 off-target sites. All detected off-target base-editing substrates contained a C within the five-base activity window (see Supplementary Information). We also monitored C→T mutations at 3,200 cytosines surrounding the six on-target and 44 off-target loci tested and observed no evident increase in C→T conversions outside the protospacer upon BE1, BE2, or BE3 treatment compared to that of untreated cells (Extended Data Fig. 7). Taken together, these findings suggest that off-target substrates of base editors include a subset of Cas9 off-target substrates, and that base editors in human cells do not induce untargeted C→T conversion throughout the genome.

Finally, we tested the potential of base editing to correct two disease-relevant mutations in mammalian cells. The apolipoprotein E gene variant *APOE4* encodes two arginine residues at amino acid positions 112 and 158, and is the largest and most common genetic risk factor for late-onset Alzheimer's disease[23]. ApoE variants with Cys

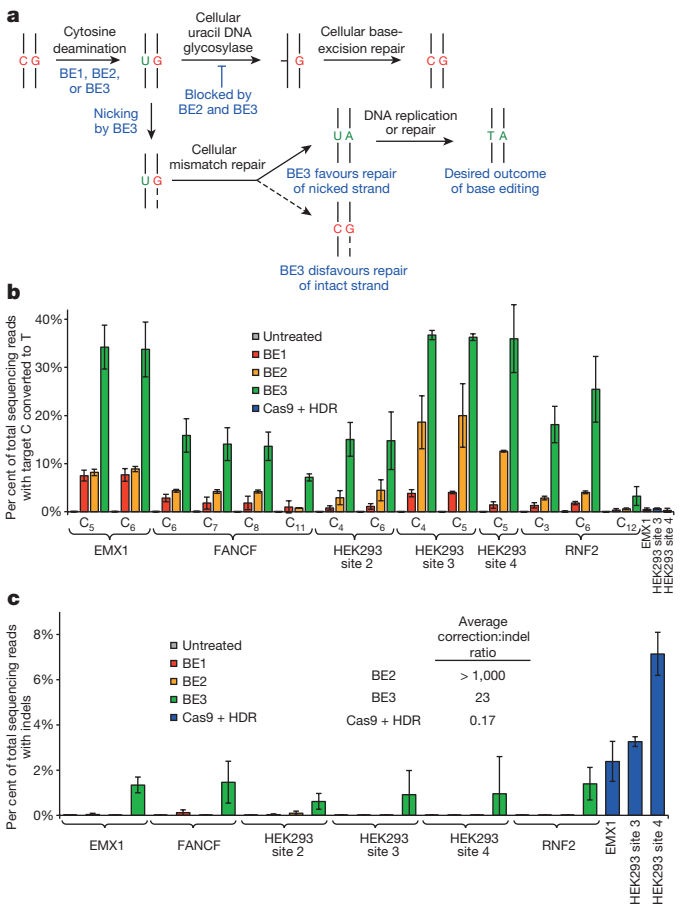


Figure 3 | Base editing in human cells. a, Possible base editing outcomes in mammalian cells. Initial editing results in a U:G mismatch. Recognition and excision of the U by uracil DNA glycosylase (UDG) initiates base excision repair (BER), which leads to reversion to the C:G starting state. BER is impeded by BE2 and BE3, which inhibit UDG. The U:G mismatch is also processed by mismatch repair (MMR), which preferentially repairs the nicked strand of a mismatch. BE3 nicks the non-edited strand containing the G, favouring resolution of the U:G mismatch to the desired U:A or T:A outcome. **b**, HEK293T cells were treated as described in the Methods. The percentage of total DNA sequencing reads with Ts at the target positions indicated are shown for treatment with BE1, BE2, or BE3, or for treatment with wild-type Cas9 with a 200-nt donor HDR template. **c**, Frequency of indel formation (see Methods) is shown following the treatment in **b**. Values are listed in Extended Data Table 1. For **b** and **c**, values and error bars reflect the mean and s.d. of three independent biological replicates performed on different days.

residues at these positions, including *APOE2* (Cys112 and Cys158), *APOE3* (Cys112 and Arg158), and *APOE3r* (Arg112 and Cys158) have been shown or are presumed[24] to confer lower Alzheimer's disease risk than *APOE4*. We attempted to convert *APOE4* into *APOE3r* in immortalized mouse astrocytes in which the endogenous *Apoe* gene was replaced by human *APOE4*. We delivered into these astrocytes by nucleofection DNA encoding BE3 and an appropriate sgRNA placing the target C at position 5 relative to a downstream PAM. After two days, we isolated nucleofected cells and measured editing efficiency by HTS of genomic DNA. We observed conversion of Arg158 to Cys158 in 58–75% of total DNA sequencing reads (Fig. 4a and Extended Data Fig. 8a). We also observed 36–50% editing of total DNA at the third position of codon 158 and 38–55% editing of total DNA at the first position of Leu159, as expected since all three of these Cs are within the base editing window. Neither of the other two C→T conversions, however, alters the amino acid sequence of the ApoE3r protein, as both TGC and TGT encode Cys (all C→T changes at the third position of a codon are silent), and both CTG and TTG encode Leu.

© 2016 Macmillan Publishers Limited. All rights reserved

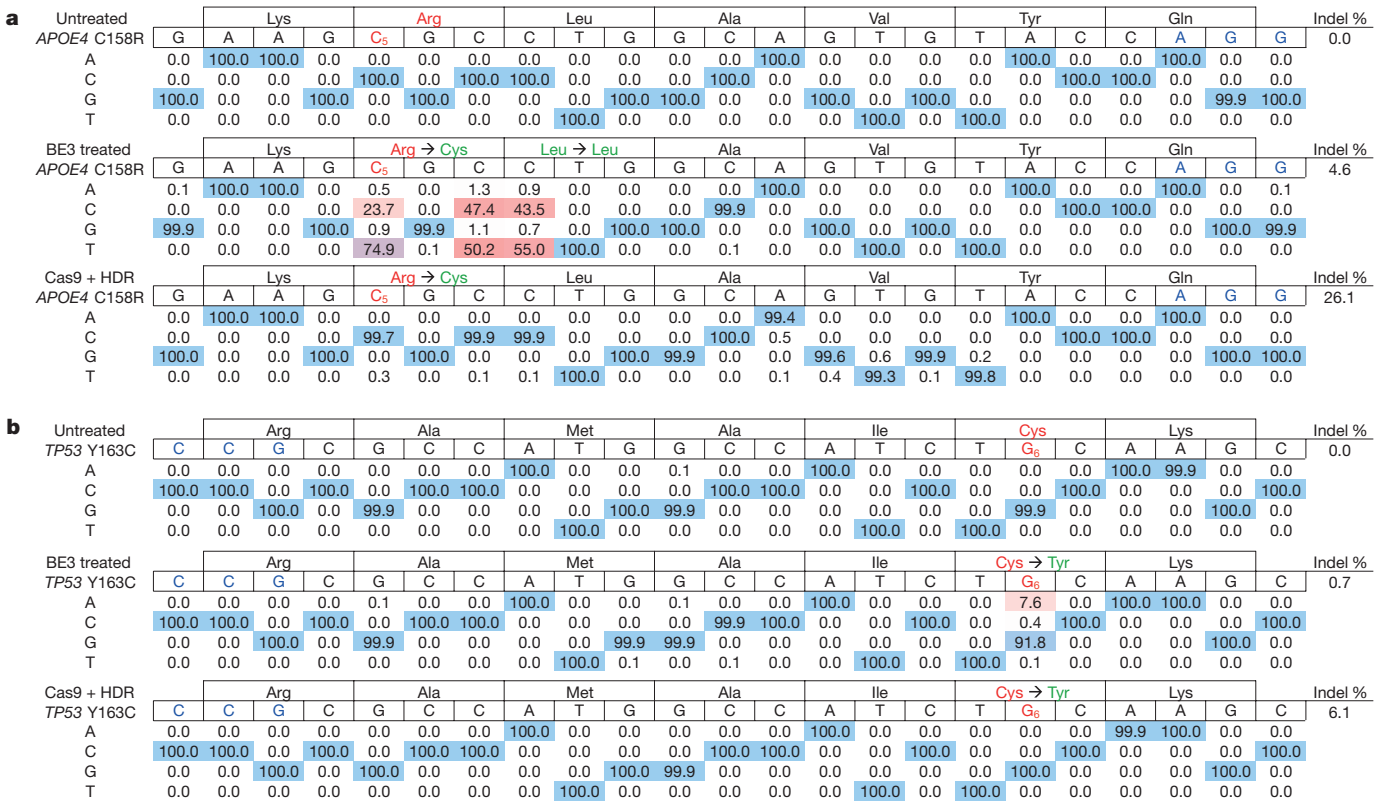


Figure 4 | BE3-mediated correction of two disease-relevant mutations in mammalian cells. The sequence of the protospacer is shown to the right of the mutation, with the PAM in blue and the target base in red with a subscripted number indicating its position within the protospacer. Underneath each sequence are the percentages of total sequencing reads with the corresponding base. Cells were treated as described in the Methods. **a**, The Alzheimer's disease-associated *APOE4* allele was converted to *APOE3r* in mouse astrocytes by BE3 in 74.9% of total reads. Two nearby Cs are also converted to Ts, but with no change to the predicted sequence of the resulting protein. Identical treatment of these cells with wild-type Cas9 and a 200-nt ssDNA donor results in 0.3% correction, with 26.1% indel formation. **b**, The cancer-associated p53 Y163C mutation is corrected by BE3 in 7.6% of nucleofected human breast cancer cells with 0.7% indel formation. Identical treatment of these cells with wild-type Cas9 and donor ssDNA results in no mutation correction with 6.1% indel formation.

The efficiency of BE3-mediated editing of *APOE4* demonstrates that a combination of suppressing BER and guiding MMR to repair the unedited strand enables base editing efficiencies to exceed the 50% maximum yield that would result from DNA replication alone. We observed no evident increase in mutations within 50 bp of either end of the protospacer compared with untreated controls (Supplementary Table 6). We observed 4.6–6.1% indels at the targeted locus following BE3 treatment. In contrast, identical treatment of astrocytes with wild-type Cas9 and donor ssDNA resulted in 0.1–0.3% *APOE4* correction and 26–40% indels at the targeted locus, efficiencies consistent with previous reports of single-base correction using Cas9 and HDR[5,6] (Fig. 4a and Extended Data Fig. 8a). Astrocytes treated identically but with an sgRNA targeting the *VEGFA* locus displayed no evidence of *APOE4* base editing (Supplementary Table 6 and Extended Data Fig. 8a). These results demonstrate that base editors can mediate highly efficient and precise single amino acid changes in the coding sequence of a protein, even when their processivity results in >1 nucleotide change in genomic DNA.

The dominant-negative p53 mutation Tyr163Cys is strongly associated with several types of cancer[25] and can be corrected by a C→T conversion on the template strand (Extended Data Fig. 2), resulting in the translation of corrected protein even before the edited base is made permanent by DNA replication or DNA repair. We nucleofected a human breast cancer cell line homozygous for the p53 Tyr163Cys mutation (HCC1954 cells) with DNA encoding BE3 and an sgRNA programmed to correct Tyr163Cys. We observed correction of the Tyr163Cys mutation in 3.3–7.6% of nucleofected HCC1954 cells (Fig. 4b, Extended Data Fig. 8b, and Supplementary Table 7), with ≤ 0.7% indel formation. In contrast, treatment of cells with wild-type Cas9 and donor ssDNA resulted in no detectable *TP53* correction (<0.1%) with 6.1–8.0% indels at the target locus (Fig. 4b and Extended Data Fig. 8b). These results collectively represent the correction of disease-associated point mutations in mammalian cell lines with an efficiency and lack of other genome modification events that may not be achievable using previously described methods. An additional 300–900 clinically relevant known human genetic diseases that in principle are correctable by the base editors described in this work are shown in Extended Data Fig. 9 and Supplementary Table 8 (see Supplementary Information).

The development of base editing advances both the scope and effectiveness of genome editing. The base editors described here offer researchers a choice of editing with very little (<0.1%) indel formation (BE2), or more efficient editing with ≤1% indel formation (BE3). That the product of base editing is, by definition, no longer a substrate likely contributes to editing efficiency by preventing subsequent product transformation, which can hamper traditional Cas9 applications. By removing the reliance on dsDNA cleavage, donor templates, and stochastic DNA repair processes that vary by cell state and cell type, base editing has the potential to expand the type of genome modifications that can be cleanly installed, the efficiency of these modifications, and the type of cells that are amenable to editing. It is likely that engineered Cas9 variants[26–28] or delivery methods[29] that offer improved DNA specificity or altered PAM specificities[30] can provide additional base editors with improved properties. These results also suggest architectures for the fusion of other DNA-modifying enzymes, including methylases and demethylases, that may enable additional types of programmable genome and epigenome base editing.

© 2016 Macmillan Publishers Limited. All rights reserved

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 26 February; accepted 30 March 2016.
Published online 20 April 2016.

1. Cox, D. B., Platt, R. J. & Zhang, F. Therapeutic genome editing: prospects and challenges. *Nature Med.* **21,** 121–131 (2015).
2. Hilton, I. B. & Gersbach, C. A. Enabling functional genomics with genome engineering. *Genome Res.* **25,** 1442–1455 (2015).
3. Sander, J. D. & Joung, J. K. CRISPR–Cas systems for editing, regulating and targeting genomes. *Nature Biotechnol.* **32,** 347–355 (2014).
4. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44,** D862–D868 (2015).
5. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339,** 819–823 (2013).
6. Ran, F. A. *et al.* Genome engineering using the CRISPR–Cas9 system. *Nature Protocols* **8,** 2281–2308 (2013).
7. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337,** 816–821 (2012).
8. Conticello, S. G. The AID/APOBEC family of nucleic acid mutators. *Genome Biol.* **9,** 229 (2008).
9. Harris, R. S., Petersen-Mahrt, S. K. & Neuberger, M. S. RNA editing enzyme APOBEC1 and some of its homologs can act as DNA mutators. *Mol. Cell* **10,** 1247–1253 (2002).
10. Jore, M. M. *et al.* Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nature Struct. Mol. Biol.* **18,** 529–536 (2011).
11. Jiang, F. *et al.* Structures of a CRISPR–Cas9 R-loop complex primed for DNA cleavage. *Science* (2016).
12. Tsai, S. Q. *et al.* Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nature Biotechnol.* **32,** 569–576 (2014).
13. Schellenberger, V. *et al.* A recombinant polypeptide extends the *in vivo* half-life of peptides and proteins in a tunable manner. *Nature Biotechnol.* **27,** 1186–1190 (2009).
14. Saraconi, G., Severi, F., Sala, C., Mattiuz, G. & Conticello, S. G. The RNA editing enzyme APOBEC1 induces somatic mutations and a compatible mutational signature is present in esophageal adenocarcinomas. *Genome Biology* **15,** 417 (2014).
15. Tsai, S. Q. *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR–Cas nucleases. *Nature Biotechnol.* **33,** 187–197 (2015).
16. Kunz, C., Saito, Y. & Schar, P. DNA repair in mammalian cells: mismatched repair: variations on a theme. *Cell. Mol. Life Sci.* **66,** 1021–1038 (2009).
17. Mol, C. D. *et al.* Crystal structure of human uracil–DNA glycosylase in complex with a protein inhibitor: protein mimicry of DNA. *Cell* **82,** 701–708 (1995).
18. Lieber, M. R., Ma, Y., Pannicke, U. & Schwarz, K. Mechanism and regulation of human non-homologous DNA end-joining. *Nature Rev. Mol. Cell Biol.* **4,** 712–720 (2003).
19. Heller, R. C. & Marians, K. J. Replisome assembly and the direct restart of stalled replication forks. *Nature Rev. Mol. Cell Biol.* **7,** 932–943 (2006).
20. Pluciennik, A. *et al.* PCNA function in the activation and strand direction of MutLα endonuclease in mismatch repair. *Proc. Natl Acad. Sci. USA* **107,** 16066–16071 (2010).
21. Beale, R. C. *et al.* Comparison of the differential context-dependence of DNA deamination by APOBEC enzymes: correlation with mutation spectra *in vivo*. *J. Mol. Biol.* **337,** 585–596 (2004).
22. Kuscu, C., Arslan, S., Singh, R., Thorpe, J. & Adli, M. Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nature Biotechnol.* **32,** 677–683 (2014).
23. Kim, J., Basak, J. M. & Holtzman, D. M. The role of apolipoprotein E in Alzheimer's disease. *Neuron* **63,** 287–303 (2009).
24. Seripa, D. *et al.* The missing ApoE allele. *Ann. Hum. Genet.* **71,** 496–500 (2007).
25. Stephens, P. J. *et al.* The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486,** 400–404 (2012).
26. Slaymaker, I. M. *et al.* Rationally engineered Cas9 nucleases with improved specificity. *Science* **351,** 84–88 (2015).
27. Davis, K. M., Pattanayak, V., Thompson, D. B., Zuris, J. A. & Liu, D. R. Small molecule-triggered Cas9 protein with improved genome-editing specificity. *Nature Chem. Biol.* **11,** 316–318 (2015).
28. Kleinstiver, B. P. *et al.* High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529,** 490–495 (2016).
29. Zuris, J. A. *et al.* Cationic lipid-mediated delivery of proteins enables efficient protein-based genome editing *in vitro* and *in vivo*. *Nature Biotechnol.* **33,** 73–80 (2015).
30. Kleinstiver, B. P. *et al.* Engineered CRISPR–Cas9 nucleases with altered PAM specificities. *Nature* **523,** 481–485 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by US National Institutes of Health (NIH) R01 EB022376 (formerly R01 GM065400), F-Prime Biomedical Research Initiative (A28161), and the Howard Hughes Medical Institute. A.C.K. is a Ruth L. Kirchstein National Research Service Awards Postdoctoral Fellow (F32 GM 112366-2). Y.B.K. holds a Natural Sciences and Engineering Research Council of Canada Postgraduate Scholarship (NSERC PGS-D). M.S.P. is an NSF Graduate Research Fellow and was supported by the Harvard Biophysics NIH training grant T32 GM008313. J.A.Z. was a Ruth L. Kirschstein National Research Service Award Postdoctoral Fellow (F32 GM 106601-2). We thank B. Hyman and E. Hudry for providing immortalized mouse astrocytes containing *APOE4*.

Author Contributions A.C.K. and Y.B.K. designed the research, performed experiments, analysed data, and wrote the manuscript. M.S.P. assisted with the data analysis. J.A.Z. assisted with the preparation of materials and the design of experiments. D.R.L. designed and supervised the research and wrote the manuscript. All of the authors contributed to editing the manuscript.

Author Information High-throughput sequencing data have been deposited in the NCBI Sequence Read Archive database under accession code SRP072434. Plasmids encoding BE1, BE2, and BE3 are available from Addgene (plasmids 73018, 73019, 73020, 73021). Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.R.L. (drliu@fas.harvard.edu).

© 2016 Macmillan Publishers Limited. All rights reserved

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The investigators were not blinded to allocation during experiments and outcome assessment.

Cloning. DNA sequences of all substrates and primers used in this paper are listed in the Supplementary Information. PCR was performed using VeraSeq ULtra DNA polymerase (Enzymatics), or Q5 Hot Start High-Fidelity DNA Polymerase (New England Biolabs). BE plasmids were constructed using USER cloning (New England Biolabs). Deaminase genes were synthesized as gBlocks Gene Fragments (Integrated DNA Technologies), and Cas9 genes were obtained from previously reported plasmids[31]. Deaminase and fusion genes were cloned into pCMV (mammalian codon-optimized) or pET28b (*E. coli* codon-optimized) backbones. sgRNA expression plasmids were constructed using site-directed mutagenesis. Briefly, the primers listed in the Supplementary Information were 5′ phosphorylated using T4 Polynucleotide Kinase (New England Biolabs) according to the manufacturer's instructions. Next, PCR was performed using Q5 Hot Start High-Fidelity Polymerase (New England Biolabs) with the phosphorylated primers and the plasmid pFYF1320 (EGFP sgRNA expression plasmid) as a template according to the manufacturer's instructions. PCR products were incubated with DpnI (20 U, New England Biolabs) at 37 °C for 1 h, purified on a QIAprep spin column (Qiagen), and ligated using QuickLigase (New England Biolabs) according to the manufacturer's instructions. DNA vector amplification was carried out using Mach1 competent cells (ThermoFisher Scientific).

***In vitro* deaminase assay on ssDNA.** Sequences of all ssDNA substrates are listed in the Supplementary Information. All Cy3-labelled substrates were obtained from Integrated DNA Technologies (IDT). Deaminases were expressed *in vitro* using the TNT T7 Quick Coupled Transcription/Translation Kit (Promega) according to the manufacturer's instructions using 1 μg of plasmid. Following protein expression, 5 μl of lysate was combined with 35 μl of ssDNA (1.8 μM) and USER enzyme (1 unit) in CutSmart buffer (New England Biolabs) (50 mM potassium acetate, 29 mM Tris-acetate, 10 mM magnesium acetate, 100 μg ml^{-1} BSA, pH 7.9) and incubated at 37 °C for 2 h. Cleaved U-containing substrates were resolved from full-length unmodified substrates on a 10% TBE-urea gel (Bio-Rad).

Expression and purification of His_6–rAPOBEC1-linker–dCas9 fusions. *E. coli* BL21 STAR (DE3)-competent cells (ThermoFisher Scientific) were transformed with plasmids encoding pET28b-His_6-rAPOBEC1-linker-dCas9 with GGS, $(GGS)_3$, XTEN, or $(GGS)_7$ linkers. The resulting expression strains were grown overnight in Luria-Bertani (LB) broth containing 100 μg ml^{-1} of kanamycin at 37 °C. The cells were diluted 1:100 into the same growth medium and grown at 37 °C to OD_{600} = ~0.6. The culture was cooled to 4 °C over a period of 2 h, and isopropyl-β-D-1-thiogalactopyranoside (IPTG) was added at 0.5 mM to induce protein expression. After ~16 h, the cells were collected by centrifugation at 4,000*g* and resuspended in lysis buffer (50 mM tris(hydroxymethyl)-aminomethane (Tris)-HCl (pH 7.5), 1 M NaCl, 20% glycerol, 10 mM tris(2-carboxyethyl)phosphine (TCEP, Soltec Ventures)). The cells were lysed by sonication (20 s pulse-on, 20 s pulse-off for 8 min total at 6 W output) and the lysate supernatant was isolated following centrifugation at 25,000*g* for 15 min. The lysate was incubated with His-Pur nickel-nitriloacetic acid (nickel-NTA) resin (ThermoFisher Scientific) at 4 °C for 1 h to capture the His-tagged fusion protein. The resin was transferred to a column and washed with 40 ml of lysis buffer. The His-tagged fusion protein was eluted in lysis buffer supplemented with 285 mM imidazole, and concentrated by ultrafiltration (Amicon-Millipore, 100-kDa molecular weight cut-off) to 1 ml total volume. The protein was diluted to 20 ml in low-salt purification buffer containing 50 mM tris(hydroxymethyl)-aminomethane (Tris)-HCl (pH 7.0), 0.1 M NaCl, 20% glycerol, 10 mM TCEP and loaded onto SP Sepharose Fast Flow resin (GE Life Sciences). The resin was washed with 40 ml of this low-salt buffer, and the protein eluted with 5 ml of activity buffer containing 50 mM tris(hydroxymethyl)-aminomethane (Tris)-HCl (pH 7.0), 0.5 M NaCl, 20% glycerol, 10 mM TCEP. The eluted proteins were quantified by SDS–PAGE.

***In vitro* transcription of sgRNAs.** Linear DNA fragments containing the T7 promoter followed by the 20-bp sgRNA target sequence were transcribed *in vitro* using the primers listed in the Supplementary Information with the TranscriptAid T7 High Yield Transcription Kit (ThermoFisher Scientific) according to the manufacturer's instructions. sgRNA products were purified using the MEGAclear Kit (ThermoFisher Scientific) according to the manufacturer's instructions and quantified by UV absorbance.

Preparation of Cy3-conjugated dsDNA substrates. Sequences of 80-nt unlabelled strands are listed in the Supplementary Information and were ordered as PAGE-purified oligonucleotides from IDT. The 25-nt Cy3-labelled primer listed in the Supplementary Information is complementary to the 3′ end of each 80-nt substrate. This primer was ordered as an HPLC-purified oligonucleotide from IDT. To generate the Cy3-labelled dsDNA substrates, the 80-nt strands (5 μl of a 100 μM solution) were combined with the Cy3-labelled primer (5 μl of a 100 μM solution) in NEBuffer 2 (38.25 μl of a 50 mM NaCl, 10 mM Tris-HCl, 10 mM $MgCl_2$, 1 mM DTT, pH 7.9 solution, New England Biolabs) with dNTPs (0.75 μl of a 100 mM solution) and heated to 95 °C for 5 min, followed by a gradual cooling to 45 °C at a rate of 0.1 °C per s. After this annealing period, Klenow exo$^-$ (5 U, New England Biolabs) was added and the reaction was incubated at 37 °C for 1 h. The solution was diluted with buffer PB (250 μl, Qiagen) and isopropanol (50 μl) and purified on a QIAprep spin column (Qiagen), eluting with 50 μl of Tris buffer.

Deaminase assay on dsDNA. The purified fusion protein (20 μl of 1.9 μM in activity buffer) was combined with 1 equivalent of appropriate sgRNA and incubated at ambient temperature for 5 min. The Cy3-labelled dsDNA substrate was added to final concentration of 125 nM and the resulting solution was incubated at 37 °C for 2 h. The dsDNA was separated from the fusion by the addition of buffer PB (100 μl, Qiagen) and isopropanol (25 μl) and purified on a EconoSpin micro spin column (Epoch Life Science), eluting with 20 μl of CutSmart buffer (New England Biolabs). USER enzyme (1 U, New England Biolabs) was added to the purified, edited dsDNA and incubated at 37 °C for 1 h. The Cy3-labelled strand was fully denatured from its complement by combining 5 μl of the reaction solution with 15 μl of a DMSO-based loading buffer (5 mM Tris, 0.5 mM EDTA, 12.5% glycerol, 0.02% bromophenol blue, 0.02% xylene cyan, 80% DMSO). The full-length C-containing substrate was separated from any cleaved, U-containing edited substrates on a 10% TBE-urea gel (Bio-Rad) and imaged on a GE Amersham Typhoon imager.

Preparation of *in vitro*-edited dsDNA for high-throughput sequencing. The oligonucleotides listed in the Supplementary Information were obtained from IDT. Complementary sequences were combined (5 μl of a 100 μM solution) in Tris buffer and annealed by heating to 95 °C for 5 min, followed by a gradual cooling to 45 °C at a rate of 0.1 °C per s to generate 60-bp dsDNA substrates. Purified fusion protein (20 μl of 1.9 μM in activity buffer) was combined with 1 equivalent of appropriate sgRNA and incubated at ambient temperature for 5 min. The 60-mer dsDNA substrate was added to final concentration of 125 nM and the resulting solution was incubated at 37 °C for 2 h. The dsDNA was separated from the fusion by the addition of buffer PB (100 μl, Qiagen) and isopropanol (25 μl) and purified on a EconoSpin micro spin column (Epoch Life Science), eluting with 20 μl of Tris buffer. The resulting edited DNA (1 μl was used as a template) was amplified by PCR using the high-throughput sequencing primer pairs specified in the Supplementary Information and VeraSeq Ultra (Enzymatics) according to the manufacturer's instructions with 13 cycles of amplification. PCR reaction products were purified using RapidTips (Diffinity Genomics), and the purified DNA was amplified by PCR with primers containing sequencing adapters, purified, and sequenced on a MiSeq high-throughput DNA sequencer (Illumina) as previously described[32].

Cell culture. HEK293T (ATCC CRL-3216) and U2OS (ATCC HTB-96) were maintained in Dulbecco's Modified Eagle's Medium plus GlutaMax (ThermoFisher) supplemented with 10% (v/v) fetal bovine serum (FBS), at 37 °C with 5% CO_2. HCC1954 cells (ATCC CRL-2338) were maintained in RPMI-1640 medium (ThermoFisher Scientific) supplemented as described above. Immortalized mouse astrocytes containing the APOE4 isoform of the *APOE* gene (Taconic Biosciences) were cultured in Dulbecco's Modified Eagle's Medium plus GlutaMax (ThermoFisher Scientific) supplemented with 10% (v/v) fetal bovine serum (FBS) and 200 μg ml^{-1} Geneticin (ThermoFisher Scientific).

Transfections. HEK293T cells were seeded on 48-well collagen-coated BioCoat plates (Corning) and transfected at approximately 85% confluency. Briefly, 750 ng of BE and 250 ng of sgRNA expression plasmids were transfected using 1.5 μl of Lipofectamine 2000 (ThermoFisher Scientific) per well according to the manufacturer's protocol.

Astrocytes, U2OS, HCC1954 and HEK293T cells were transfected using appropriate Amaxa Nucleofector II programs according to manufacturer's instructions (basic glial cell, V, V, and V kits using programs T-020, X-001, X-005, and Q-001 for astrocytes, U2OS, HCC1954, and HEK293T cells, respectively). 40 ng of iRFP670 (Addgene plasmid 45457)[33] was added to the nucleofection solution to assess nucleofection efficiencies in these cell lines. Astrocytes and HCC1954 cells were filtered through a 40 μm strainer (Fisher Scientific) after harvesting, and the nucleofected cells were collected on a Beckman Coulter MoFlo XDP Cell Sorter using the iRFP signal (absorbance 643 nm, emission 670 nm). The U2OS and HEK293T cells were used without enrichment of nucleofected cells.

High-throughput DNA sequencing of genomic DNA samples. Transfected cells were harvested after 3 days and the genomic DNA was isolated using the Agencourt DNAdvance Genomic DNA Isolation Kit (Beckman Coulter) according to the manufacturer's instructions. On-target and off-target genomic regions of interest were amplified by PCR with flanking high-throughput sequencing primer pairs listed in the Supplementary Information. PCR amplification was carried out with

© 2016 Macmillan Publishers Limited. All rights reserved

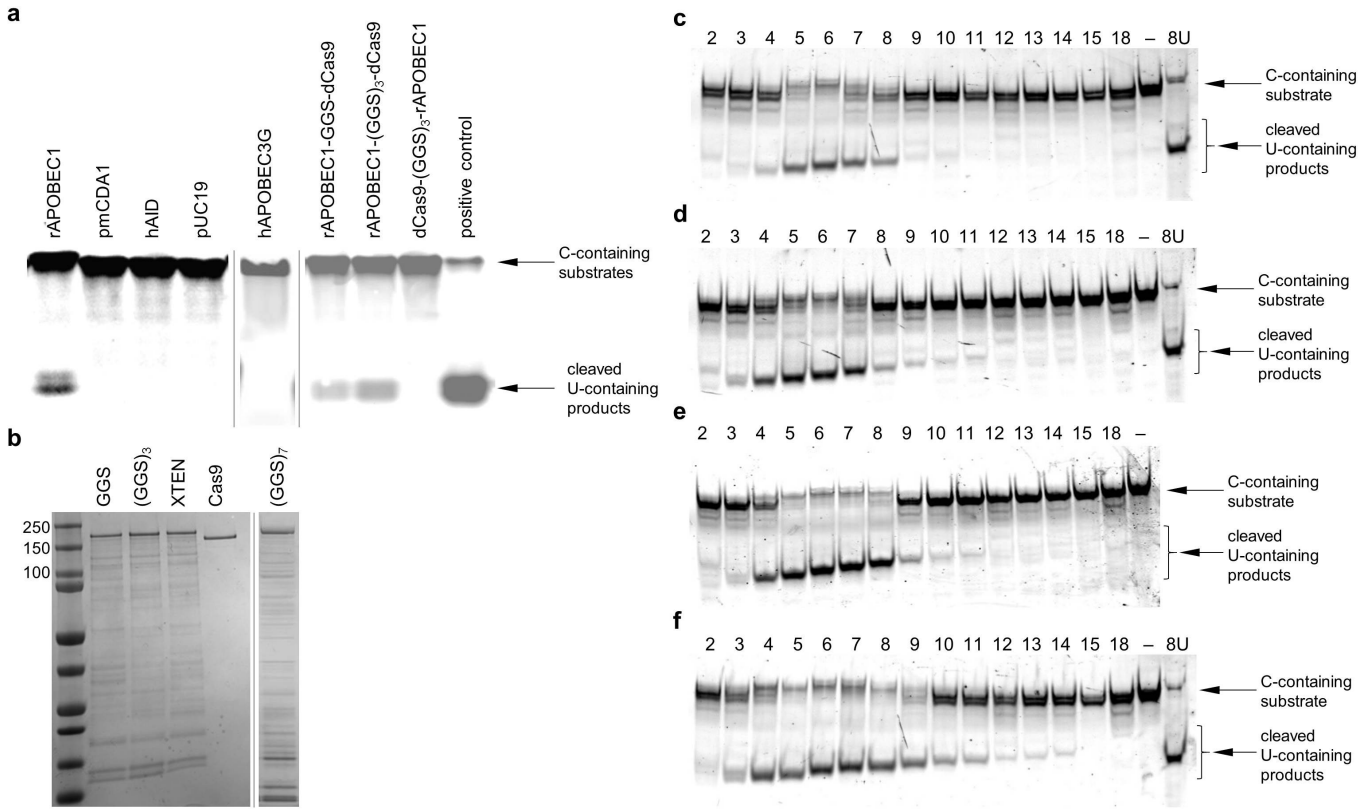
Phusion high-fidelity DNA polymerase (ThermoFisher) according to the manufacturer's instructions using 5 ng of genomic DNA as a template. Cycle numbers were determined separately for each primer pair as to ensure the reaction was stopped in the linear range of amplification (30, 28, 28, 28, 32, and 32 cycles for EMX1, FANCF, HEK293 site 2, HEK293 site 3, HEK293 site 4, and RNF2 primers, respectively). PCR products were purified using RapidTips (Diffinity Genomics). Purified DNA was amplified by PCR with primers containing sequencing adaptors. The products were gel purified and quantified using the Quant-iT PicoGreen dsDNA Assay Kit (ThermoFisher) and KAPA Library Quantification Kit-Illumina (KAPA Biosystems). Samples were sequenced on an Illumina MiSeq as previously described[32].

Data analysis. Sequencing reads were automatically demultiplexed using MiSeq Reporter (Illumina), and individual FASTQ files were analysed with a custom Matlab script provided in the Supplementary Information. Each read was pairwise aligned to the appropriate reference sequence using the Smith-Waterman algorithm. Base calls with a Q-score below 31 were replaced with Ns and were thus excluded in calculating nucleotide frequencies. This treatment yields an expected MiSeq base-calling error rate of approximately 1 in 1,000. Aligned sequences in which the read and reference sequence contained no gaps were stored in an alignment table from which base frequencies could be tabulated for each locus.

Indel frequencies were quantified with a custom Matlab script shown in the Supplementary Information using previously described criteria[29]. Sequencing reads were scanned for exact matches to two 10-bp sequences that flank both sides of a window in which indels might occur. If no exact matches were located, the read was excluded from analysis. If the length of this indel window exactly matched the reference sequence the read was classified as not containing an indel. If the indel window was two or more bases longer or shorter than the reference sequence, then the sequencing read was classified as an insertion or deletion, respectively.

31. Mali, P. *et al.* CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nature Biotechnol.* **31,** 833–838 (2013).
32. Pattanayak, V. *et al.* High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nature Biotechnol.* **31,** 839–843 (2013).
33. Shcherbakova, D. M. & Verkhusha, V. V. Near-infrared fluorescent proteins for multicolor in vivo imaging. *Nature Methods* **10,** 751–754 (2013).

© 2016 Macmillan Publishers Limited. All rights reserved



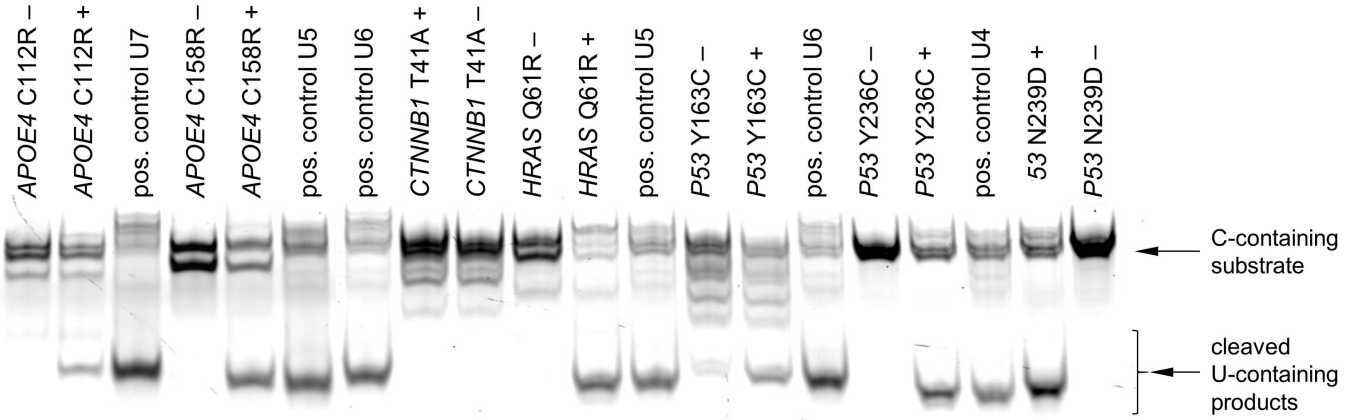
Extended Data Figure 1 | Effects of deaminase, linker length, and linker composition on base editing. a, Gel-based deaminase assay showing activity of rat APOBEC1 (rAPOBEC1), lamprey CDA1 (pmCDA1), human AID (hAID), human APOBEC3G (hAPOBEC3G), rAPOBEC1-GGS-dCas9, rAPOBEC1-$(GGS)_3$-dCas9, and dCas9-$(GGS)_3$-rAPOBEC1 on ssDNA. Enzymes were expressed in a mammalian cell lysate-derived *in vitro* transcription-translation system and incubated with 1.8 μM dye-conjugated ssDNA and USER enzyme (uracil DNA glycosylase and endonuclease VIII) at 37 °C for 2 h. The resulting DNA was resolved on a denaturing polyacrylamide gel and imaged. The positive control is a sequence with a U synthetically incorporated at the same position as the target C. **b**, Coomassie-stained denaturing PAGE of the expressed and purified proteins used in **c**–**f**. **c**–**f**, Gel-based deaminase assay showing the deamination window of base editors with deaminase–Cas9 linkers of GGS (**c**), $(GGS)_3$ (**d**), XTEN (**e**), or $(GGS)_7$ (**f**). Following incubation of 1.85 μM deaminase–dCas9 fusions complexed with sgRNA with 125 nM dsDNA substrates at 37 °C for 2 h, the dye-conjugated DNA was isolated and incubated with USER enzyme at 37 °C for 1 h to cleave the DNA backbone at the site of any Us. The resulting DNA was resolved on a denaturing polyacrylamide gel, and the dye-conjugated strand was imaged. Each lane is numbered according to the position of the target C within the protospacer, or labelled with '–' if no target C is present. 8U is a positive control sequence with a U synthetically incorporated at position 8. For uncropped gel data, see Supplementary Fig. 1.

© 2016 Macmillan Publishers Limited. All rights reserved

a

APOE4 Cys112Arg:	5'-GGAGGACGTGC_{11}GCGGCCGCCTGG
APOE4 Cys158Arg:	5'-GAAGC_5GCCTGGCAGTGTACCAGG
CTNNB1 Thr41Ala:	5'-CTGTGGC_7AGTGGCACCAGAATGG
HRAS Gln61Arg:	5'-CCTCCC_6GGCCGGCGGTATCCAGG
P53 Tyr163Cys:	5'-GCTTGC_6AGATGGCCATGGCGCGG
P53 Tyr236Cys:	5'-ACACATGC_8AGTTGTAGTGGATGG
P53 Asn239Asp:	5'-TGTC_4ACACATGTAGTTGTAGTGG

b



Extended Data Figure 2 | BE1 is capable of correcting disease-relevant mutations *in vitro*. a, Protospacer and PAM sequences (blue) of seven disease-relevant mutations. The disease-associated target C in each case is indicated with a subscripted number reflecting its position within the protospacer. For all mutations except both *APOE4* SNPs, the target C resides in the template (non-coding) strand. **b**, Deaminase assay showing each dsDNA 80-mer oligonucleotide before (–) and after (+) incubation with BE1, DNA isolation, and incubation with USER enzymes to cleave DNA at positions containing U. Positive control lanes from incubation of synthetic oligonucleotides containing U at various positions within the protospacer with USER enzymes are shown with the corresponding number indicating the position of the U. Editing efficiencies were quantitated by dividing the intensity of the cleaved product band by that of the entire lane for each sample. For uncropped gel data, see Supplementary Fig. 1.

© 2016 Macmillan Publishers Limited. All rights reserved

protospacer and PAM sequence: 5'-TTCCCCCCCCGATTTATTTATGG-3'

sequence	% of total reads
...CCCCCCCC...	62.4
...TTTTTTCC...	18.2
...TTTTTTTC...	13.4
...TTTTTTTT...	3.3
...TCCCCCCC...	0.8
...CCCCTTCC...	0.3
...CCCTTTCC...	0.3
...TTTTTCCC...	0.3
...CCCCTCCC...	0.3

Extended Data Figure 3 | Processivity of BE1. The protospacer and PAM (blue) of a 60-mer DNA oligonucleotide containing eight consecutive Cs is shown at the top. The oligonucleotide (125 nM) was incubated with BE1 (2 μM) for 2 h at 37 °C. The DNA was isolated and analysed by high-throughput sequencing. Shown are the percent of total reads for the most frequent nine sequences observed. The vast majority of edited strands (>93%) have more than one C converted to T.

© 2016 Macmillan Publishers Limited. All rights reserved

a

EMX1: GAGTC$_5$C$_6$GAGCAGAAGAAGAAGGG
FANCF: GGAATC$_6$C$_7$C$_8$TTC$_{11}$TGCAGCACCTGG
HEK293 site 2: GAAC$_4$AC$_6$AAAGCATAGACTGCGGG
HEK293 site 3: GGCC$_4$C$_5$AGACTGAGCACGTGATGG
HEK293 site 4: GGCAC$_5$TGCGGCTGGAGGTCCGGG
RNF2: GTC$_3$ATC$_6$TTAGTC$_{12}$ATTACCTGAGG

b

EMX1		G	A	G	T	C$_5$	C$_6$	G	A	G	C	A	G	A	A	G	A	A	G	A	A	G	G	G
1.85 uM BE1	A	0.2	99.9	0.1	0.1	0.1	0.0	0.3	100.0	0.1	0.1	99.9	0.2	99.9	99.7	0.3	100.0	99.9	0.2	99.9	97.9	0.1	0.2	0.2
	C	0.0	0.0	0.0	0.0	60.4	61.0	0.0	0.0	0.0	99.1	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0
	G	99.7	0.0	99.9	0.0	0.0	0.0	99.7	0.0	99.9	0.1	0.0	99.7	0.0	0.0	99.7	0.0	0.0	99.8	0.0	0.0	99.8	99.8	99.8
	T	0.1	0.0	0.0	99.9	39.5	39.0	0.0	0.0	0.0	0.7	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

FANCF		G	G	A	A	T	C$_6$	C$_7$	C$_8$	T	T	C$_{11}$	T	G	C	A	G	C	A	C	C	T	G	G
1.85 uM BE1	A	0.1	0.2	99.8	99.9	0.0	0.1	0.1	0.1	0.0	0.1	0.1	0.1	0.1	0.1	99.9	0.2	0.2	95.0	0.1	0.1	0.1	0.3	0.4
	C	0.0	0.0	0.1	0.1	0.0	63.9	64.7	65.0	0.0	0.0	72.6	0.0	0.0	99.4	0.1	0.0	99.7	4.9	99.8	99.9	0.0	0.0	0.0
	G	99.8	99.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	99.8	0.2	0.0	99.7	0.2	0.0	0.1	0.0	0.0	99.6	99.5
	T	0.1	0.1	0.0	0.0	99.9	36.0	35.1	34.9	99.9	99.9	27.3	99.8	0.1	0.3	0.0	0.1	0.0	0.0	0.0	0.0	99.9	0.1	0.1

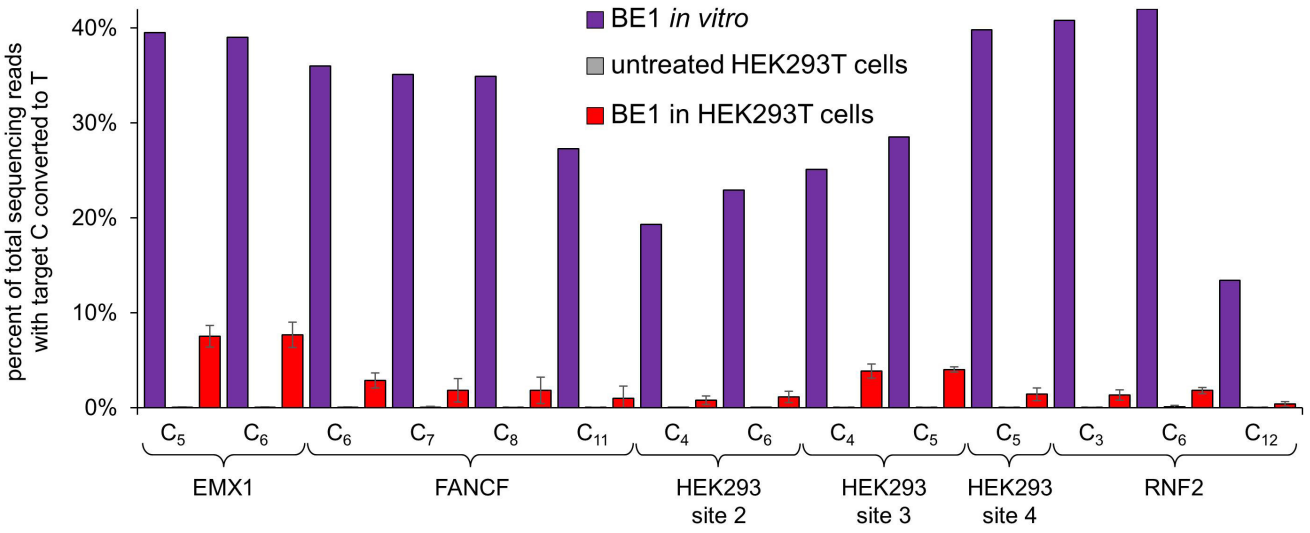
HEK293 site 2		G	A	A	C$_4$	A	C$_6$	A	A	A	G	C	A	T	A	G	A	C	T	G	C	G	G	G
1.85 uM BE1	A	0.1	100.0	99.8	0.1	100.0	0.1	99.9	100.0	99.9	0.1	0.1	100.0	0.0	99.9	0.2	100.0	0.1	0.0	0.1	0.1	0.1	0.2	0.2
	C	0.0	0.0	0.1	80.6	0.0	76.9	0.0	0.0	0.0	0.0	99.6	0.0	0.0	0.1	0.0	0.0	99.9	0.0	0.0	99.9	0.0	0.0	0.0
	G	99.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	99.8	0.0	0.0	0.1	0.0	99.8	0.0	0.0	0.0	99.8	0.0	99.9	99.8	99.8
	T	0.1	0.0	0.0	19.3	0.0	22.9	0.0	0.0	0.0	0.1	0.3	0.0	99.9	0.0	0.1	0.0	0.0	99.9	0.0	0.1	0.0	0.1	0.0

HEK293 site 3		G	G	C	C$_4$	C$_5$	A	G	A	C	T	G	A	G	C	A	C	G	T	G	A	T	G	G
1.85 uM BE1	A	0.2	0.1	0.1	0.1	0.0	99.9	0.2	99.9	0.1	0.0	0.2	99.9	0.1	0.1	100.0	0.1	0.1	0.0	0.2	98.6	0.0	0.1	0.1
	C	0.0	0.0	92.2	74.8	71.5	0.0	0.0	0.0	96.6	0.0	0.0	0.0	0.0	99.8	0.0	99.7	0.0	0.0	0.0	1.3	0.0	0.0	0.0
	G	99.7	99.8	0.0	0.0	0.0	0.0	99.8	0.0	0.0	0.0	99.8	0.0	99.8	0.0	0.0	0.0	99.8	0.0	99.8	0.0	0.0	99.9	99.8
	T	0.1	0.1	7.7	25.1	28.5	0.0	0.0	0.0	3.3	99.9	0.1	0.0	0.0	0.1	0.0	0.2	0.0	99.9	0.0	0.0	100.0	0.0	0.0

HEK293 site 4		G	G	C	A	C$_5$	T	G	C	G	G	C	T	G	G	A	G	G	T	G	G	G	G	G
1.85 uM BE1	A	0.3	0.2	0.1	100.0	0.1	0.0	0.2	0.1	0.2	0.2	0.1	0.0	0.3	0.5	100.0	0.1	0.1	0.0	0.1	0.1	0.1	0.2	0.2
	C	0.0	0.0	98.8	0.0	60.1	0.0	0.0	97.0	0.0	0.0	99.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	G	99.7	99.7	0.0	0.0	0.0	0.0	99.8	0.0	99.8	99.8	0.0	0.0	99.7	99.4	0.0	99.8	99.8	0.1	99.8	99.8	99.8	99.8	99.7
	T	0.0	0.1	1.1	0.0	39.8	100.0	0.1	2.9	0.0	0.1	0.5	100.0	0.0	0.1	0.0	0.0	0.0	99.9	0.0	0.1	0.1	0.0	0.0

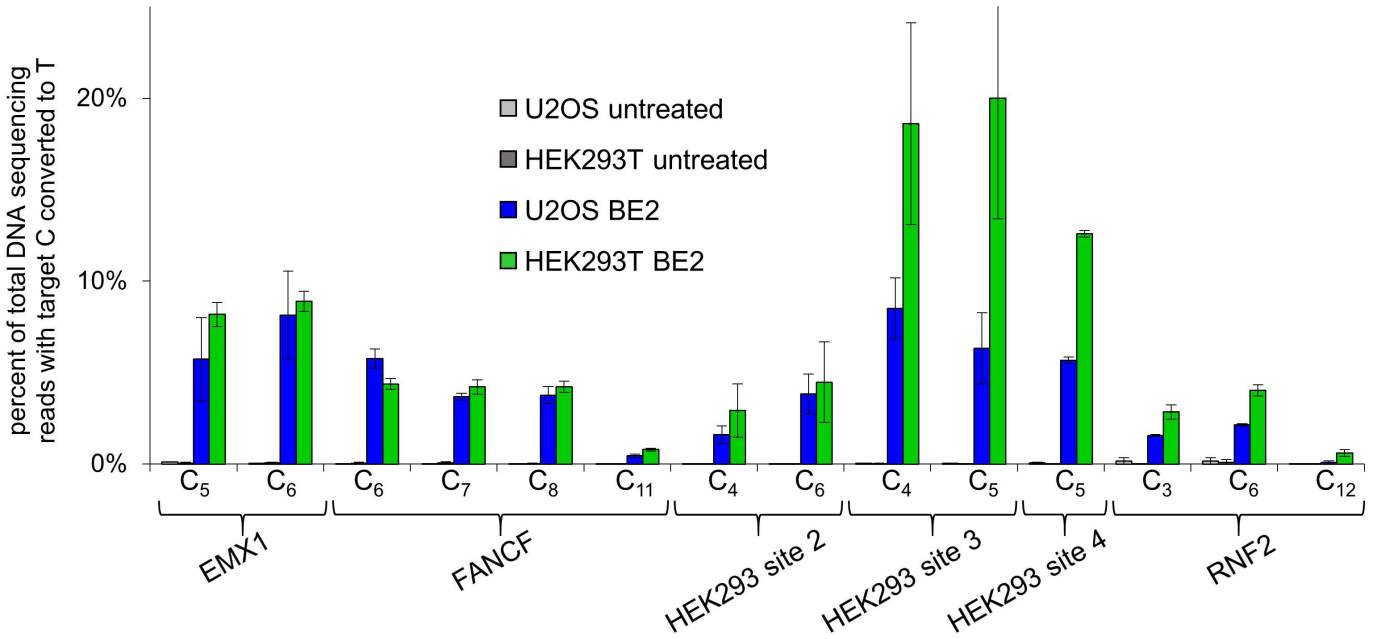
RNF2		G	T	C$_3$	A	T	C$_6$	T	T	A	G	T	C$_{12}$	A	T	T	A	C	C	T	G	A	G	G
1.85 uM BE1	A	0.1	0.0	0.1	99.9	0.0	0.0	0.0	0.0	99.9	0.1	0.0	0.1	99.9	0.0	0.0	99.9	0.1	0.0	0.0	0.1	98.6	0.0	0.0
	C	0.0	0.0	59.1	0.0	0.0	57.8	0.0	0.0	0.1	0.0	0.0	86.4	0.0	0.0	0.0	0.1	99.9	99.9	0.0	0.0	1.4	0.0	0.0
	G	99.9	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	99.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	99.8	0.0	99.9	99.9
	T	0.0	99.9	40.8	0.0	100.0	42.1	100.0	100.0	0.0	0.1	99.9	13.4	0.0	100.0	99.9	0.0	0.0	0.0	99.9	0.1	0.0	0.1	0.0

c



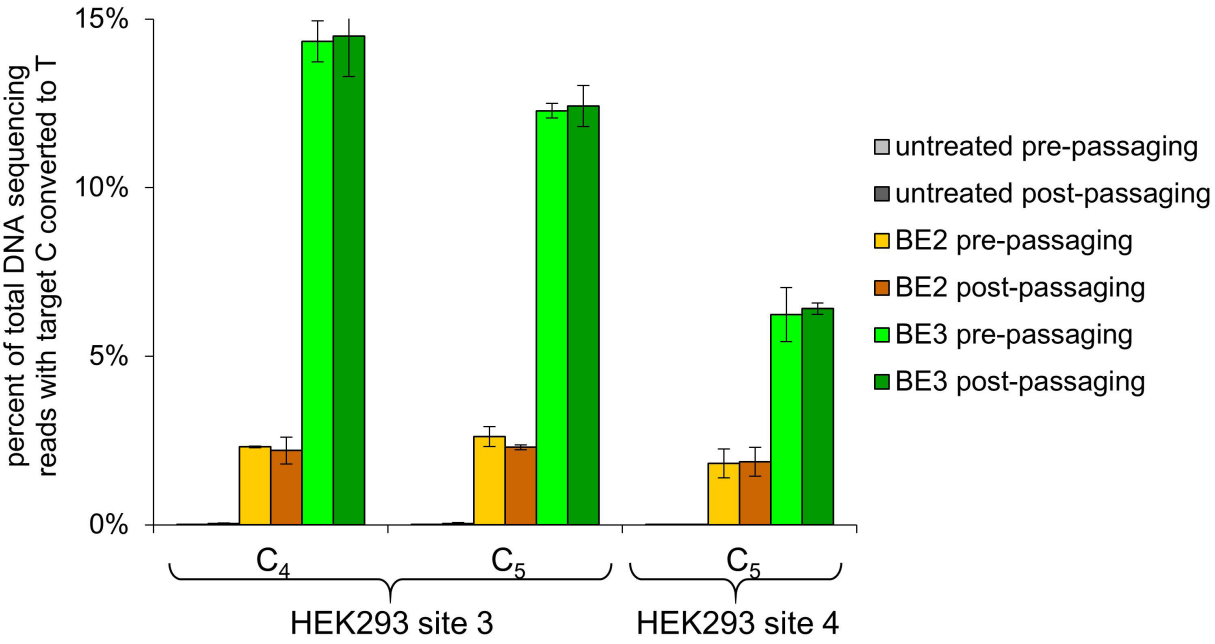
Extended Data Figure 4 | BE1 base editing efficiencies are strikingly decreased in mammalian cells. a, Protospacer (black and red) and PAM (blue) sequences of the six mammalian cell genomic loci targeted by base editors. Target Cs are indicated in red with subscripted numbers corresponding to their positions within the protospacer. **b**, Synthetic 80-mers with sequences matching six different genomic sites were incubated with BE1 then analysed for base editing by high-throughput sequencing. For each site, the sequence of the protospacer is indicated to the right of the name of the site, with the PAM highlighted in blue. Underneath each sequence are the percentages of total DNA sequencing reads with the corresponding base. We considered a target C as 'editable' if the *in vitro* conversion efficiency is >10%. Note that maximum yields are 50% of total DNA sequencing reads since the non-targeted strand is unaffected by BE1. Values are shown from a single experiment. **c**, HEK293T cells were transfected with plasmids expressing BE1 and an appropriate sgRNA. Three days after transfection, genomic DNA was extracted and analysed by high-throughput sequencing at the six loci. Cellular C to T conversion percentages, defined as the percentage of total DNA sequencing reads with Ts at the target positions indicated, are shown for BE1 at all six genomic loci. Values and error bars of all data from HEK293T cells reflect the mean and standard deviation of three independent biological replicates performed on different days.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 5 | Base editing efficiencies of BE2 in U2OS and HEK293T cells. Cellular C to T conversion percentages by BE2 are shown for each of the six targeted genomic loci in HEK293T cells and U2OS cells. HEK293T cells were transfected using Lipofectamine 2000, and U2OS cells were nucleofected. Three days after plasmid delivery, genomic DNA was extracted and analysed for base editing at the six genomic loci by HTS. Values and error bars reflect the mean and standard deviation of two (U2OS) or three (HEK293T) biological experiments performed on different days.

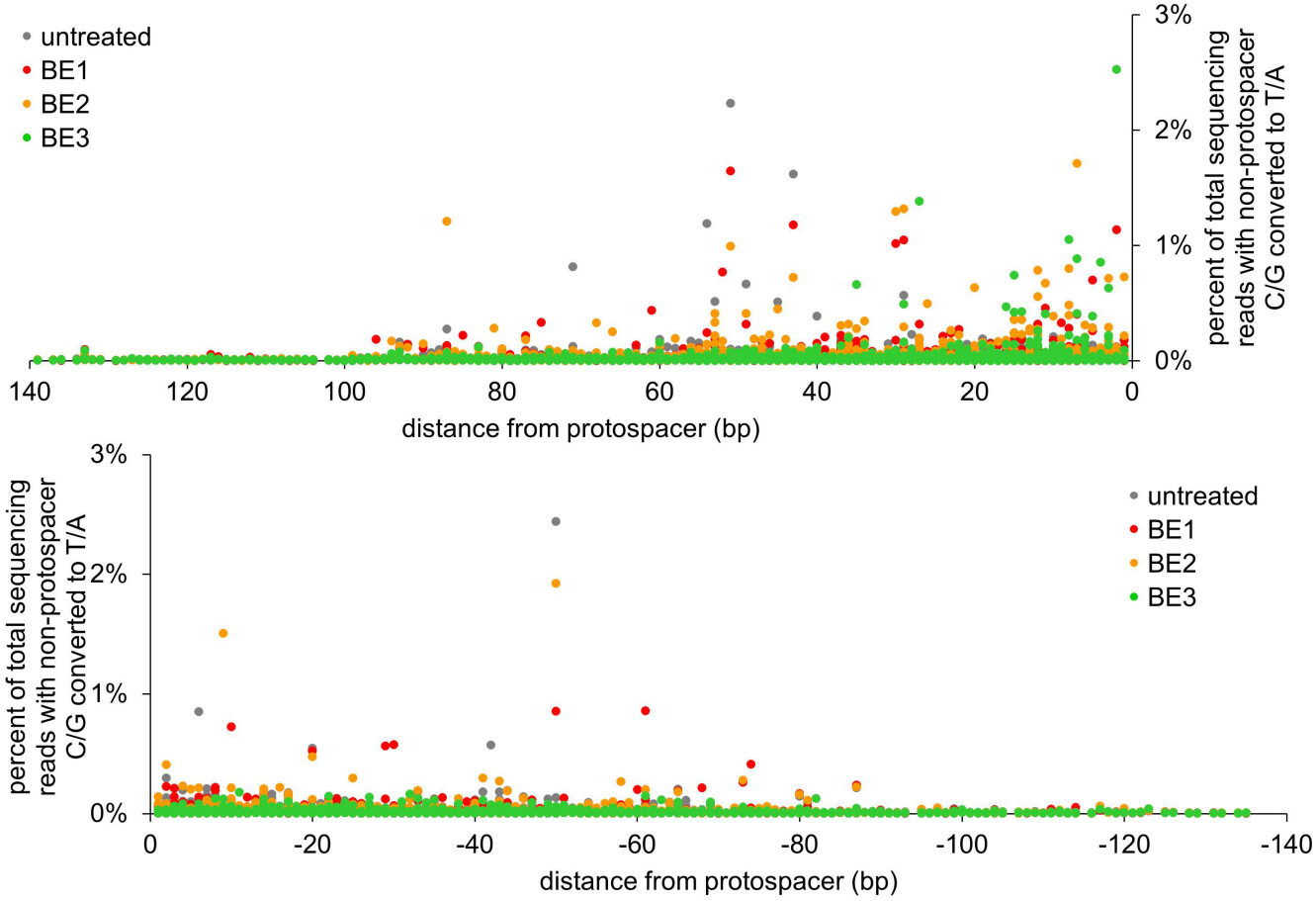
© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 6 | Base editing persists over multiple cell divisions. Cellular C to T conversion percentages by BE2 and BE3 are shown for HEK293 sites 3 and 4 in HEK293T cells before and after passaging the cells. HEK293T cells were nucleofected with plasmids expressing BE2 or BE3 and an sgRNA targeting HEK293 site 3 or 4. Three days after nucleofection, the cells were harvested and split in half. One half was subjected to high-throughput sequencing analysis, and the other half was allowed to propagate for approximately five cell divisions, then harvested and subjected to high-throughput sequencing analysis. Values and error bars reflect the mean and standard deviation of two biological experiments performed on different days.

© 2016 Macmillan Publishers Limited. All rights reserved

a



b

non-protospacer C/Gs	average C/G (%)	average T/A (%)	lowest T/A (%)	highest T/A (%)
untreated	99.95 ± 0.14	0.02 ± 0.02	0.00	2.44
BE1	99.95 ± 0.24	0.03 ± 0.03	0.00	1.64
BE2	99.95 ± 0.13	0.03 ± 0.03	0.00	1.92
BE3	99.97 ± 0.09	0.02 ± 0.02	0.00	2.52

Extended Data Figure 7 | Non-target C/G mutation rates. Shown here are the C to T and G to A mutation rates at 3,200 distinct cytosines and guanines surrounding the six on-target and 34 off-target loci tested, representing a total of 14,700,000 sequence reads derived from approximately 1.8×10^6 cells. **a**, Cellular non-target C to T and G to A conversion percentages by BE1, BE2, and BE3 are plotted individually against their positions relative to a protospacer for all 3,200 cytosines/guanines. The side of the protospacer distal to the PAM is designated with positive numbers, while the side that includes the PAM is designated with negative numbers. **b**, Average non-target cellular C to T and G to A conversion percentages by BE1, BE2, and BE3 are shown, as well as the highest and lowest individual conversion percentages.

© 2016 Macmillan Publishers Limited. All rights reserved

a

Untreated	Lys			Arg			Leu			Ala			Val			Tyr			Gln			indel %
APOE4 C158R	G	A	A	G	C$_5$	G	C	C	T	G	G	C	A	G	T	G	T	A	C	C	A	G	G	0.0
A	0.0	100.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0	
C	0.0	0.0	0.0	0.0	100.0	0.0	100.0	100.0	0.0	0.0	0.1	100.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	0.0	0.0	
G	100.0	0.0	0.0	99.9	0.0	100.0	0.0	0.0	0.0	100.0	99.9	0.0	0.0	100.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	99.9	100.0	
T	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	100.0	0.0	0.0	0.0	0.0	0.1	0.0	

BE3 + on-target sgRNA	Lys			Arg → Cys			Leu → Leu			Ala			Val			Tyr			Gln			indel %
APOE4 C158R	G	A	A	G	C$_5$	G	C	C	T	G	G	C	A	G	T	G	T	A	C	C	A	G	G	6.1
A	0.1	100.0	100.0	0.0	1.0	0.0	1.6	0.6	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.1	
C	0.0	0.0	0.0	0.0	39.1	0.0	62.0	61.3	0.0	0.0	0.0	99.9	0.0	0.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	0.0	0.0	
G	99.9	0.0	0.0	100.0	1.5	99.9	0.7	0.5	0.0	100.0	100.0	0.0	0.0	100.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	100.0	99.9	
T	0.0	0.0	0.0	0.0	58.3	0.1	35.7	37.5	100.0	0.0	0.0	0.1	0.0	0.0	100.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	

BE3 + off-target sgRNA	Lys			Arg			Leu			Ala			Val			Tyr			Gln			indel %
APOE4 C158R	G	A	A	G	C$_5$	G	C	C	T	G	G	C	A	G	T	G	T	A	C	C	A	G	G	0.0
A	0.0	100.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0	
C	0.0	0.0	0.0	0.0	99.9	0.0	100.0	100.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	0.0	0.0	
G	100.0	0.0	0.0	100.0	0.0	100.0	0.0	0.0	0.0	100.0	100.0	0.0	0.0	100.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	100.0	100.0	
T	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	

Cas9 + HDR	Lys			Arg → Cys			Leu			Ala			Val			Tyr			Gln			indel %
APOE4 C158R	G	A	A	G	C$_5$	G	C	C	T	G	G	C	A	G	T	G	T	A	C	C	A	G	G	40.1
A	0.0	99.7	99.7	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	98.0	0.0	0.0	0.0	0.0	99.1	0.1	0.5	98.8	0.0	0.0	
C	0.0	0.3	0.2	0.0	99.8	0.0	99.8	99.9	0.1	0.1	0.2	99.9	1.8	0.4	0.2	0.3	0.2	0.5	99.7	99.6	0.7	0.0	0.0	
G	99.7	0.0	0.0	99.9	0.0	99.7	0.1	0.0	0.1	99.8	99.7	0.0	0.1	98.6	1.4	99.4	0.2	0.3	0.1	0.0	0.0	99.4	100.0	
T	0.2	0.0	0.1	0.1	0.1	0.3	0.1	0.1	99.8	0.2	0.1	0.1	0.1	1.1	98.4	0.4	99.7	0.2	0.2	0.0	0.5	0.6	0.0	

b

Untreated	Arg			Ala			Met			Ala			Ile			Cys			Lys			indel %
TP53 Y163C	C	C	G	C	G	C	C	A	T	G	G	C	C	A	T	C	T	G$_6$	C	A	A	G	C	0.0
A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.1	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	100.0	99.9	0.0	0.0	
C	100.0	100.0	0.0	100.0	0.0	100.0	100.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0	0.0	100.0	
G	0.0	0.0	100.0	0.0	99.9	0.0	0.0	0.0	0.0	100.0	99.9	0.0	0.0	0.0	0.0	0.0	0.0	99.9	0.0	0.0	0.0	100.0	0.0	
T	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	

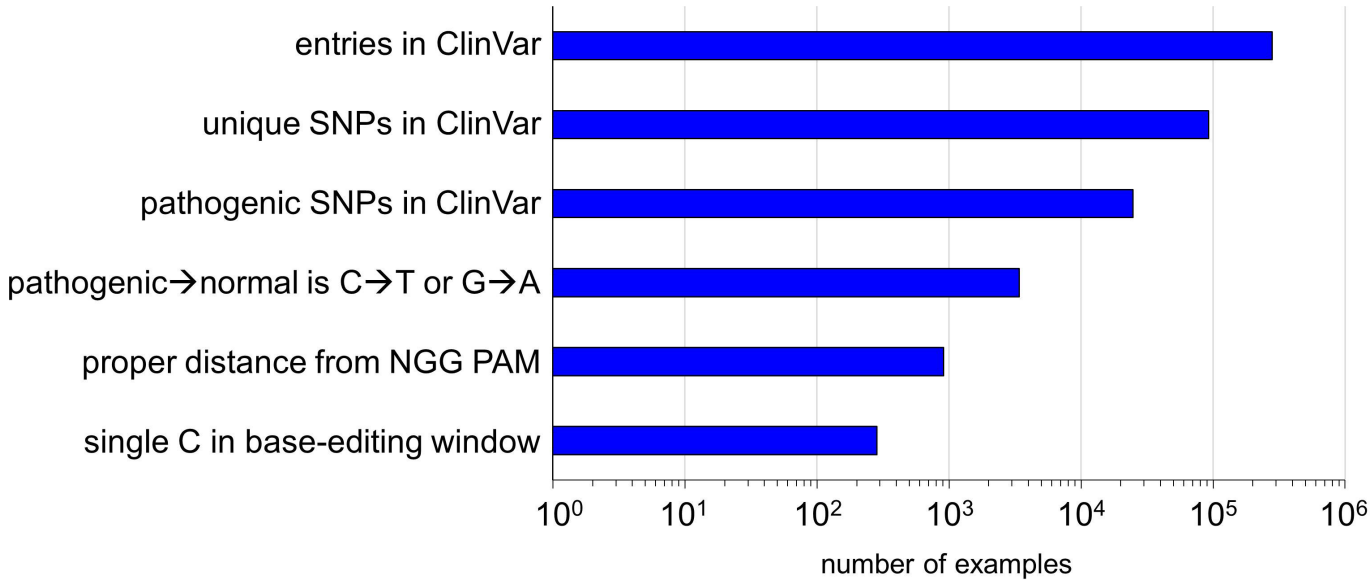
BE3 + on-target sgRNA	Arg			Ala			Met			Ala			Ile			Cys → Tyr			Lys			indel %
TP53 Y163C	C	C	G	C	G	C	C	A	T	G	G	C	C	A	T	C	T	G$_6$	C	A	A	G	C	0.0
A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	99.9	0.0	0.0	0.1	0.0	0.0	100.0	0.0	0.0	0.0	3.3	0.5	99.9	99.9	0.0	0.0	
C	100.0	100.0	0.0	100.0	0.0	100.0	100.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	0.0	100.0	0.0	0.0	99.5	0.0	0.0	0.0	100.0	
G	0.0	0.0	100.0	0.0	99.9	0.0	0.0	0.0	0.0	100.0	99.9	0.0	0.0	0.0	0.0	0.0	0.0	96.2	0.0	0.0	0.0	100.0	0.0	
T	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	100.0	0.5	0.0	0.0	0.0	0.0	0.0	

BE3 + off-target sgRNA	Arg			Ala			Met			Ala			Ile			Cys			Lys			indel %
TP53 Y163C	C	C	G	C	G	C	C	A	T	G	G	C	C	A	T	C	T	G$_6$	C	A	A	G	C	0.0
A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	99.9	0.0	0.0	0.1	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	99.9	99.9	0.0	0.0	
C	100.0	100.0	0.0	99.9	0.0	100.0	100.0	0.0	0.0	0.0	0.0	100.0	99.9	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0	0.0	100.0	
G	0.0	0.0	100.0	0.0	99.9	0.0	0.0	0.0	0.0	100.0	99.9	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	100.0	0.0	
T	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.1	0.0	100.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	

Cas9 + HDR	Arg			Ala			Met			Ala			Ile			Cys → Tyr			Lys			indel %
TP53 Y163C	C	C	G	C	G	C	C	A	T	G	G	C	C	A	T	C	T	G$_6$	C	A	A	G	C	8.0
A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.1	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	99.9	100.0	0.0	0.4	
C	100.0	100.0	0.0	100.0	0.0	100.0	100.0	0.0	0.0	0.0	0.0	100.0	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0	0.0	99.6	
G	0.0	0.0	100.0	0.0	99.9	0.0	0.0	0.0	0.0	100.0	99.9	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	100.0	0.0	
T	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	

Extended Data Figure 8 | Additional data sets of BE3-mediated correction of two disease-relevant mutations in mammalian cells. For each site, the sequence of the protospacer is indicated to the right of the name of the mutation, with the PAM highlighted in blue and the base responsible for the mutation indicated in red bold with a subscripted number corresponding to its position within the protospacer. The amino acid sequence above each disease-associated allele is shown, together with the corrected amino acid sequence following base editing in green. Underneath each sequence are the percentages of total sequencing reads with the corresponding base. Cells were nucleofected with plasmids encoding BE3 and an appropriate sgRNA. Two days after nucleofection, genomic DNA was extracted from the nucleofected cells and analysed by high-throughput sequencing to assess pathogenic mutation correction. **a**, The Alzheimer's disease-associated *APOE4* allele is converted to *APOE3r* in mouse astrocytes by BE3 in 58.3% of total reads only when treated with the correct sgRNA. Two nearby Cs are also converted to Ts, but with no change to the predicted sequence of the resulting protein. Identical treatment of these cells with wild-type Cas9 and a 200-nt ssDNA donor results in 0.2% correction, with 26.7% indel formation. **b**, The cancer-associated p53 Y163C mutation is corrected by BE3 in 3.3% of nucleofected human breast cancer cells only when treated with the correct sgRNA. Identical treatment of these cells with wild-type Cas9 and donor ssDNA results in no detectable mutation correction with 8.0% indel formation.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 9 | Genetic variants from ClinVar that, in principle, can be corrected by base editing. The NCBI ClinVar database of human genetic variations and their corresponding phenotypes (see main text ref. 4) was searched for genetic diseases that can be corrected by current base editing technologies. The results were filtered by imposing the successive restrictions listed on the left. The *x* axis shows the number of occurrences satisfying that restriction and all above restrictions on a logarithmic scale.

© 2016 Macmillan Publishers Limited. All rights reserved

Extended Data Table 1 | Indel formation following treatment of HEK293T cells with BE1, BE2, BE3, or wild-type Cas9 plus a ssDNA template for HDR

	EMX1 indel (%)	FANCF indel (%)	HEK293 site 2 indel (%)	HEK293 site 3 indel (%)	HEK293 site 4 indel (%)	RNF2 indel (%)
untreated	0.01±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00	0.00±0.00
BE1	0.04±0.05	0.11±0.13	0.02±0.04	0.00±0.00	0.00±0.00	0.00±0.00
BE2	0.01±0.00	0.01±0.01	0.09±0.09	0.00±0.00	0.00±0.00	0.00±0.00
BE3	1.34±0.35	1.47±0.93	0.62±0.35	0.91±1.07	0.95±1.64	1.39±0.72
Cas9 + HDR	2.38±0.89			3.26±0.22	7.14±0.96	

Indel frequencies were calculated as described in the Methods following treatment of HEK293T cells with BE1, BE2, and BE3 for all six genomic loci, or with wild-type Cas9 and a ssDNA template for HDR at three of the six sites (EMX1, HEK293 site 3, and HEK293 site 4). Values reflect the mean and standard deviation of at least three independent biological replicates performed on different days.

© 2016 Macmillan Publishers Limited. All rights reserved

LETTER

doi:10.1038/nature17994

Structure of the thermally stable Zika virus

Victor A. Kostyuchenko[1,2]*, Elisa X. Y. Lim[1,2]*, Shuijun Zhang[1,2]*, Guntur Fibriansah[1,2], Thiam-Seng Ng[1,2], Justin S. G. Ooi[1,2], Jian Shi[2,3] & Shee-Mei Lok[1,2]

Zika virus (ZIKV), formerly a neglected pathogen, has recently been associated with microcephaly in fetuses[1], and with Guillian–Barré syndrome in adults[2]. Here we present the 3.7 Å resolution cryo-electron microscopy structure of ZIKV, and show that the overall architecture of the virus is similar to that of other flaviviruses. Sequence and structural comparisons of the ZIKV envelope (E) protein with other flaviviruses show that parts of the E protein closely resemble the neurovirulent West Nile and Japanese encephalitis viruses, while others are similar to dengue virus (DENV). However, the contribution of the E protein to flavivirus pathobiology is currently not understood. The virus particle was observed to be structurally stable even when incubated at 40 °C, in sharp contrast to the less thermally stable DENV[3]. This is also reflected in the infectivity of ZIKV compared to DENV serotypes 2 and 4 (DENV2 and DENV4) at different temperatures. The cryo-electron microscopy structure shows a virus with a more compact surface. This structural stability of the virus may help it to survive in the harsh conditions of semen[4], saliva[5] and urine[6]. Antibodies or drugs that destabilize the structure may help to reduce the disease outcome or limit the spread of the virus.

Zika virus (ZIKV), a flavivirus, is thought to be principally transmitted to humans by the mosquito (*Aedes aegypti*) vector. Other flaviviruses include West Nile virus (WNV), Japanese encephalitis virus (JEV), dengue virus (DENV) and yellow fever virus (YFV). ZIKV generally causes a mild disease. However, when pregnant women are infected with ZIKV, there is an increased risk of developing microcephaly in the fetus[1].

Retrospective analysis of data collected from a ZIKV outbreak in French Polynesia in 2013–2014 showed association of the virus with microcephaly[7]. Here we present the 3.7 Å resolution structure of ZIKV strain H/PF/2013 isolated during that outbreak[8].

For cryo-electron microscopy (cryoEM) studies, ZIKV was grown in the mosquito cell line at 28 °C and purified at 4 °C by polyethylene glycol precipitation, a sucrose cushion, followed by a potassium tartrate gradient. The gel analysis of the purified sample suggested it contained mostly mature virus (Extended Data Fig. 1). The ZIKV samples were incubated at 28 °C, 37 °C and 40 °C (mimicking high fever) for 30 min, before imaging by cryoEM (Fig. 1a). At 28 °C, there were broken and shrivelled particles together with some smooth surfaced particles (about 500 Å in diameter), similar to the compact DENV mature particles. Conversely, samples incubated at 37 °C and 40 °C showed many more smooth surfaced particles. The presence of a larger fraction of shrivelled particles at 28 °C could be due to the exposure of particles to high osmolality during purification. We speculate that ZIKV particles may expand into smooth surfaced particles when incubated at higher temperatures, making the lipid envelope more fluid, and allowing the structure to revert to its normal state. Some strains of DENV2 (New Guinea C, NGC) undergo notable structural changes from smooth to bumpy surfaced particles when incubated at 37 °C (refs 9, 10), whereas others (PVP94/07)[11] do not. Structurally, DENV4 is considered more stable than DENV2 (NGC strain), as it remains as a smooth surfaced particle even at 40 °C, although considerable aggregation of particles occurs[3]. ZIKV seemed to be even more stable than DENV4, as the smooth surfaced particles did not aggregate at 40 °C (Fig. 1a). This is consistent with the plaque assay results with ZIKV, DENV2 strains (NGC and PVP94/07) and DENV4 that were pre-incubated at 28 °C, 37 °C and 40 °C for 30 min. ZIKV retained its infectivity, whereas the infectivity of DENV2 strains and DENV4 is significantly reduced at higher temperatures (Fig. 1b). The infectivity of ZIKV was only slightly reduced when the pre-incubation time was increased to 60 min (Fig. 1c).

The cryoEM structure was reconstructed from the virus sample incubated at 40 °C for 30 min. The 3.7 Å resolution cryoEM map (Fig. 2a and Extended Data Fig. 2) showed well-resolved densities corresponding to the α-helices and β-strands. In addition, many side-chain densities were visible (Fig. 2b). The overall structural architecture of ZIKV (Fig. 2c) is similar to that of DENV (refs 3, 12–14) and WNV (ref. 15). While our paper was under review, a similar ZIKV structure was published[16] showing similar features. There are 180 copies of each of the E and associated M ('membrane') proteins on the virus surface; they are anchored in the bilayer lipid membrane via their transmembrane

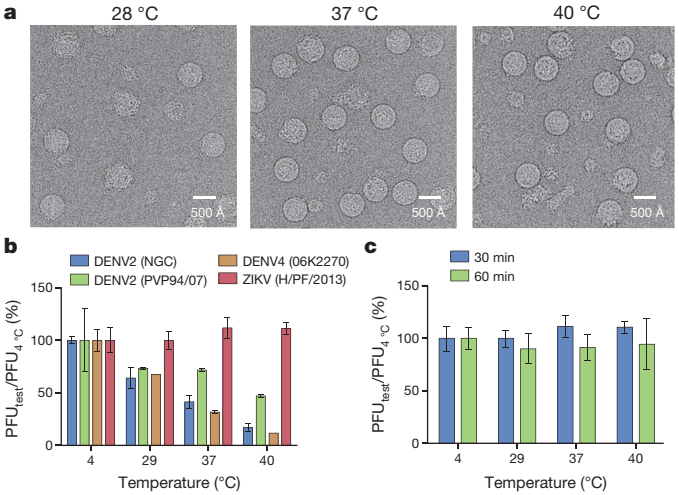


Figure 1 | ZIKV particles are more stable than DENV. a, CryoEM micrographs of ZIKV particles incubated at different temperatures. **b**, ZIKV infection in BHK-21 cells is not affected when the virus is pre-incubated for 30 min at different temperatures. By contrast, all DENV strains have reduced infectivity with increasing temperatures. The *y* axis shows the percentage of plaque-forming units (PFU) at different temperatures divided by that of the same virus at 4 °C. **c**, Increased incubation time slightly reduced ZIKV infectivity. Blue and green bars indicate 30- and 60-min incubation times, respectively. Two experiments were done, and each condition was tested in triplicate. Data are mean and s.d.; means were calculated by averaging the triplicates within an experiment, and the s.d. shows the variability between the means from two separate experiments.

[1]Program in Emerging Infectious Diseases, Duke–National University of Singapore Medical School, Singapore 169857, Singapore. [2]Centre for BioImaging Sciences, Department of Biological Sciences, National University of Singapore, Singapore 117557, Singapore. [3]CryoEM unit, Department of Biological Sciences, National University of Singapore, Singapore 117557, Singapore. *These authors contributed equally to this work.

© 2016 Macmillan Publishers Limited. All rights reserved

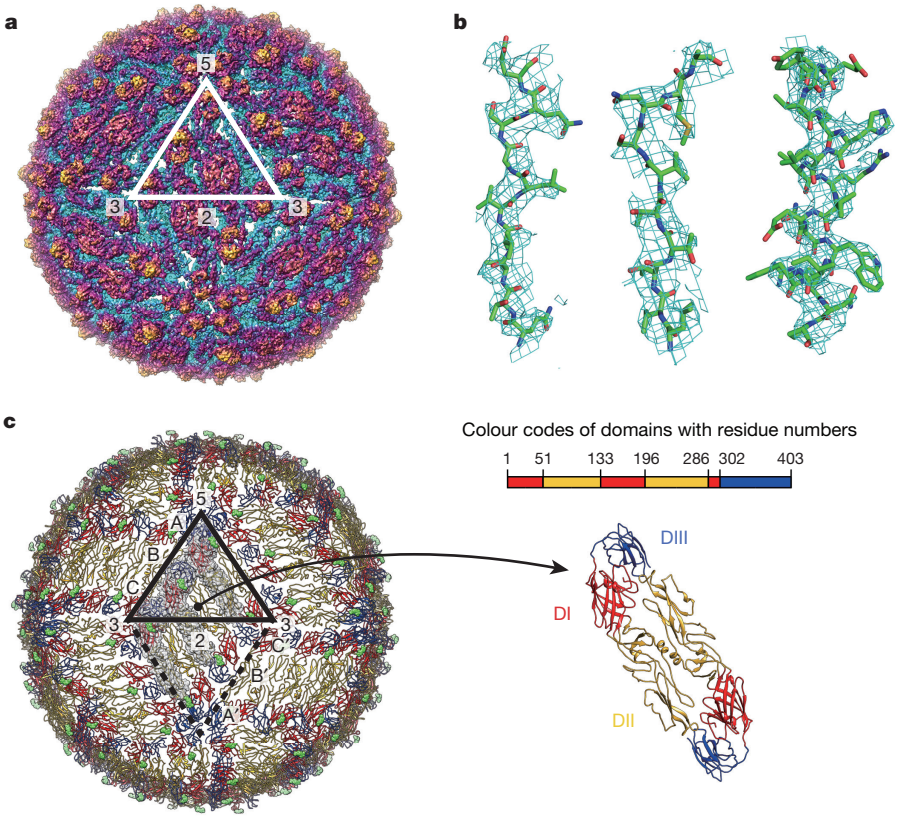


Figure 2 | The 3.7 Å resolution cryoEM structure of ZIKV. a, The ZIKV cryoEM map. The map is coloured radially: 0–219 Å (cyan), 220–240 Å (purple) and >240 Å (gold). **b**, Zoomed-in view of the density map (cyan mesh) fitted with the E protein structure. **c**, The E protein arrangement on the ZIKV surface. The Asn154 glycosylation sites are shown as green spheres. The three individual E proteins (A, B and C) in one asymmetric unit (black triangle) are indicated by transparent surfaces.

regions. The E protein is the major protein involved in receptor binding and fusion. The M protein is a small protein that is hidden under the E protein layer. The E and M proteins are organized in icosahedral symmetry consisting of 60 repeating units, and each asymmetric unit contains three individual E proteins (molecules A, B and C) (Fig. 2c). The E proteins are arranged as dimers, with three dimers lying parallel to each other forming a raft; there are 30 such rafts covering the viral surface. The E ectodomain contains three domains: DI, DII and DIII (Fig. 2c, right). DI acts as a bridge between DII and DIII. The tip of DII contains the fusion loop, which interacts with the endosomal membrane during fusion. The DI–DII hinge is thought to be important for flipping out DII to expose the fusion loop during the fusion event. The DI–DIII hinge is thought to be rigid at neutral pH. In most other flaviviruses, DIII contains the receptor-binding site[17] and has an important role in fusion[18]. After the carboxy-terminal end of the E ectodomain is the α-helical stem region that lies on the viral membrane, and the α-helical transmembrane region.

The superposition of the three individual E protein molecules (A, B and C) in the asymmetric unit of ZIKV showed high similarity (Extended Data Fig. 3a), typical for flaviviruses (Extended Data Fig. 3b). Superposition of DI of molecule A of the ZIKV E protein with that of DENV2 (ref. 14) showed a slight difference in the angle at the DI–DIII and DI–DII hinges (Extended Data Fig. 3c).

Sequence comparison shows that the ZIKV E protein has similar sequence homology levels (ranging from 40 to 58%) to other flaviviruses (Extended Data Fig. 4a). There are some insertions, deletions and mutations on the E proteins that are distinct among the neurovirulent viruses (WNV, JEV and tick-borne encephalitis virus (TBEV)) and the febrile-illness-causing viruses (YFV and DENV) (Extended Data Fig. 4b). However, their contributions to the resultant pathology are not clear. A superposition of our ZIKV cryoEM structure onto the DENV2 cryoEM[14] structure, and the crystal structures of WNV[19] and JEV[20] E proteins show structural similarities (Extended Data Fig. 5), which are consistent with the sequence comparisons (Extended Data Fig. 4b). For example, ZIKV is similar to DENV at the hi-loop of DII (Extended Data Figs 4b and 5a), where both have a three-residue long insertion compared to WNV, TBEV and JEV. However, at other sites, ZIKV is more similar to the neurovirulent JEV and WNV (Extended Data Figs 4b and 5b). For example, sequence comparison shows an Asn67Asp mutation in JEV, WNV and ZIKV compared to DENV, indicating that the glycosylation site is absent (Fig. 2c). Therefore, there is only one Asn154 glycosylation site in an E protein in ZIKV, whereas DENV has two (Asn67 and Asn153)[14]. The Asn154 glycosylation site is suggested to be a neurovirulence determinant in WNV[21]; however, DENV also has this glycosylation site but infection with DENV rarely results in neurovirulence. The Asn67 (but not Asn153) glycosylation site in DENV has been shown to be crucial for binding to DC-SIGN (also known as CD209)[22] on dendritic cells. ZIKV, which also interacts with DC-SIGN[23], does not have this glycosylation site, and therefore could have a different binding mechanism.

A longer glycan loop at DI is observed in ZIKV, WNV and JEV compared with TBEV, YFV and DENV1–4 (Extended Data Figs 4b and 5b). An extra residue is present in the DII kl-loop of WNV, TBEV, YFV and ZIKV compared to DENV1–4. A residue insertion at the DIII C strand in WNV, JEV, TBEV and ZIKV was observed as opposed to DENV1–4. Finally, there is a two-residue insertion at the DIII DE-loop in WNV, JEV and ZIKV compared to TBEV, YFV and DENV1–4 (Extended Data Figs 4b and 5b).

Notably, our ZIKV structure shows that the one-residue insertion at the end of the C strand of DIII (Extended Data Fig. 4b) compared to DENV may increase the overall stability of the virus. For comparison, a similar resolution DENV2 cryoEM structure[14], which has been shown

© 2016 Macmillan Publishers Limited. All rights reserved

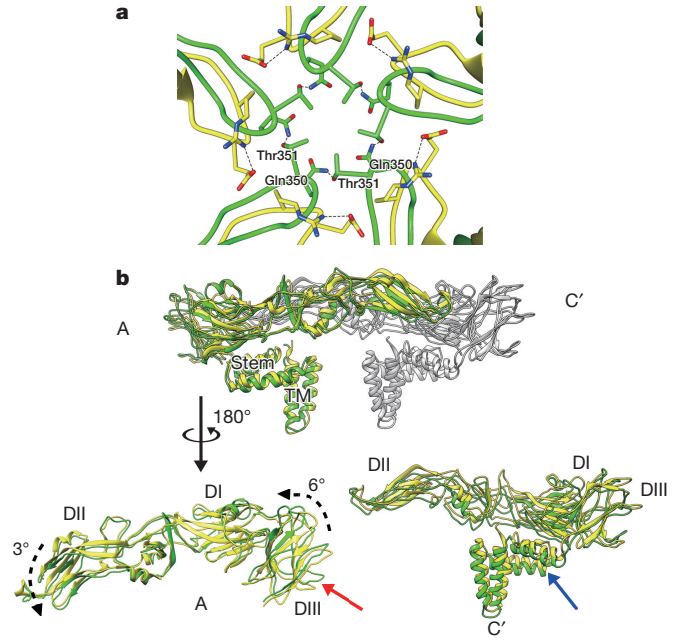


Figure 3 | Possible interactions between the ZIKV DIII CD-loops around the five-fold vertex may cause differences in the position of A–C′ dimer compared to DENV2. a, The CD-loops of ZIKV (green) are stretched closer towards the five-fold vertex than those in DENV2 (yellow). In ZIKV, the Gln350 residue may form a hydrogen bond (dashed line) to the Thr351 from the CD-loop of a neighbouring DIII, creating a hydrogen bond network around the five-fold vertex. The DENV2 residues in the CD-loop (yellow), in comparison, do not interact with other neighbouring CD-loops. **b**, Positional differences of ZIKV A–C′ dimer compared to DENV2. Top panel shows the side view. ZIKV and DENV molecule A are coloured green and yellow, respectively. Molecule C′ of both ZIKV and DENV2 is coloured grey. Bottom left panel show molecule A rotated by 180° to an axis perpendicular to the viral membrane. The ZIKV CD-loop (red arrow) is lifted upwards probably owing to the five-fold interaction as demonstrated in **a**; this molecule rotates slightly. Bottom right panel shows molecule C′ in the same orientation as in the top panel. The whole ZIKV E protein (green) molecule C′ is positioned at a lower radius than in DENV2 (yellow). The positions of the ZIKV stem (blue arrow) and transmembrane (TM) regions suggest that the membrane is pushed downwards compared to DENV2.

to be less stable at increased temperatures[9,10], was used. The Ala340 insertion in the C strand of DIII allows the CD-loop to stretch further towards the five-fold vertex compared to DENV2 (Fig. 3a). Gln350 of the CD-loop from one DIII could potentially form a hydrogen bond with Thr351 from another CD-loop in the neighbouring DIII, and thus may create a hydrogen bond network between the five DIIIs (Fig. 3a and Extended Data Fig. 6). In comparison, in DENV2, there is a lack of interaction between the DIIIs in this region (Fig. 3a). The possible hydrogen bond network around the five-fold vertex of ZIKV could be responsible for the rotation of the E protein molecule A slightly (6°) anticlockwise around the axis parallel to the viral membrane (Fig. 3b, bottom left). When compared to DENV2, the position of DIII of the ZIKV molecule A was at a slightly higher radius, whereas DII dips slightly closer to the viral membrane. The stem and transmembrane regions showed no notable positional differences in molecule A between DENV2 and ZIKV (Fig. 3b, top). The ectodomain of molecule C′, which forms a dimer with molecule A, is positioned at a lower radius than DENV2. The stem and transmembrane regions (Fig. 3b, bottom right), along with the membrane, are also pressed downward. The ectodomains of the B–B′ dimer also lie at a slightly lower radius (Extended Data Fig. 7b), although no positional difference in the stem and transmembrane regions was observed when compared to DENV2 (Extended Data Fig. 7a, bottom). These findings suggest that the ZIKV has a more compact surface than DENV2.

The E proteins on the surface of some DENV2 strains are shown to undergo structural rearrangements at 37 °C (ref. 9). In this structure, the A–C′ dimer moves outwards to a higher radius and also rotates slightly. The B–B′ dimer moves to a higher radius and the E proteins dissociate from each other. A study of an antibody, which traps a transitional stage of this motion[11], suggests that the A–C′ dimer probably moves to a higher radius before the B–B′ dimer[11]. In the ZIKV structure, the possible network of interactions around the five-fold vertex between the five A–C′ dimers may result in tighter packing that prevents this structural transition.

Both the TIM (T cell, immunoglobulin and mucin) and TAM (Tyro3, Axl and Mer) phosphatidylserine transmembrane receptors have been shown to be used by DENV for cell entry[24]. Conversely, ZIKV was shown to bind to TAM but not TIM receptors[23]. It is possible that the looser surface of DENV2 compared to ZIKV could allow more interactions with a wider range of phosphatidylserine-binding receptors.

Unlike DENV, ZIKV is detected in the urine[6], saliva[5] and semen[4] (which allows potential transmission by sexual contact[25]), suggesting that the virus is stable enough to withstand harsh conditions. This is consistent with our infection assay showing that ZIKV is more thermally stable than DENV. Furthermore, the potential interactions identified in our structure may explain the stability of ZIKV and may serve as potential sites to exploit for future therapeutics.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 24 March; accepted 11 April 2016.
Published online 19 April 2016.

1. Schuler-Faccini, L. *et al.* Possible association between Zika virus infection and microcephaly – Brazil, 2015. *MMWR Morb. Mortal. Wkly. Rep.* **65,** 59–62 (2016).
2. Cao-Lormeau, V. M. *et al.* Guillain–Barre syndrome outbreak associated with Zika virus infection in French Polynesia: a case-control study. *Lancet* **387,** 1531–1539 (2016).
3. Kostyuchenko, V. A., Chew, P. L., Ng, T. S. & Lok, S. M. Near-atomic resolution cryo-electron microscopic structure of dengue serotype 4 virus. *J. Virol.* **88,** 477–482 (2014).
4. Mansuy, J. M. *et al.* Zika virus: high infectious viral load in semen, a new sexually transmitted pathogen? *Lancet Infect. Dis.* **16,** 405 (2016).
5. Barzon, L. *et al.* Isolation of infectious Zika virus from saliva and prolonged viral RNA shedding in a traveller returning from the Dominican Republic to Italy, January 2016. *Euro Surveill.* **21,** 1 (2016).
6. Gourinat, A. C., O'Connor, O., Calvez, E., Goarant, C. & Dupont-Rouzeyrol, M. Detection of Zika virus in urine. *Emerg. Infect. Dis.* **21,** 84–86 (2015).
7. Cauchemez, S. *et al.* Association between Zika virus and microcephaly in French Polynesia, 2013–15: a retrospective study. *Lancet* http://dx.doi.org/10.1016/S0140-6736(16)00651-6 (2016).
8. Baronti, C. *et al.* Complete coding sequence of Zika virus from a French polynesia outbreak in 2013. *Genome Announc.* **2,** e00500-14 (2014).
9. Fibriansah, G. *et al.* Structural changes in dengue virus when exposed to a temperature of 37 degrees C. *J. Virol.* **87,** 7585–7592 (2013).
10. Zhang, X. *et al.* Dengue structure differs at the temperatures of its human and mosquito hosts. *Proc. Natl Acad. Sci. USA* **110,** 6795–6799 (2013).
11. Fibriansah, G. *et al.* DENGUE VIRUS. Cryo-EM structure of an antibody that neutralizes dengue virus type 2 by locking E protein dimers. *Science* **349,** 88–91 (2015).
12. Kostyuchenko, V. A., Zhang, Q., Tan, J. L., Ng, T. S. & Lok, S. M. Immature and mature dengue serotype 1 virus structures provide insight into the maturation process. *J. Virol.* **87,** 7700–7707 (2013).
13. Kuhn, R. J. *et al.* Structure of dengue virus: implications for flavivirus organization, maturation, and fusion. *Cell* **108,** 717–725 (2002).
14. Zhang, X. *et al.* Cryo-EM structure of the mature dengue virus at 3.5-Å resolution. *Nature Struct. Mol. Biol.* **20,** 105–110 (2013).
15. Mukhopadhyay, S., Kim, B. S., Chipman, P. R., Rossmann, M. G. & Kuhn, R. J. Structure of West Nile virus. *Science* **302,** 248 (2003).
16. Sirohi, D. *et al.* The 3.8 Å resolution cryo-EM structure of Zika virus. *Science* http://dx.doi.org/10.1126/science.aaf5316 (2016).
17. Rey, F. A., Heinz, F. X., Mandl, C., Kunz, C. & Harrison, S. C. The envelope glycoprotein from tick-borne encephalitis virus at 2 Å resolution. *Nature* **375,** 291–298 (1995).
18. Bressanelli, S. *et al.* Structure of a flavivirus envelope glycoprotein in its low-pH-induced membrane fusion conformation. *EMBO J.* **23,** 728–738 (2004).

© 2016 Macmillan Publishers Limited. All rights reserved

19. Nybakken, G. E., Nelson, C. A., Chen, B. R., Diamond, M. S. & Fremont, D. H. Crystal structure of the West Nile virus envelope glycoprotein. *J. Virol.* **80,** 11467–11474 (2006).
20. Luca, V. C., AbiMansour, J., Nelson, C. A. & Fremont, D. H. Crystal structure of the Japanese encephalitis virus envelope protein. *J. Virol.* **86,** 2337–2346 (2012).
21. Beasley, D. W. *et al.* Envelope protein glycosylation status influences mouse neuroinvasion phenotype of genetic lineage 1 West Nile virus strains. *J. Virol.* **79,** 8339–8347 (2005).
22. Pokidysheva, E. *et al.* Cryo-EM reconstruction of dengue virus in complex with the carbohydrate recognition domain of DC-SIGN. *Cell* **124,** 485–493 (2006).
23. Hamel, R. *et al.* Biology of Zika virus infection in human skin cells. *J. Virol.* **89,** 8880–8896 (2015).
24. Meertens, L. *et al.* The TIM and TAM families of phosphatidylserine receptors mediate dengue virus entry. *Cell Host Microbe* **12,** 544–557 (2012).
25. McCarthy, M. US health officials investigate sexually transmitted Zika virus infections. *Br. Med. J.* **352,** i1180 (2016).

Acknowledgements We thank the European Virus Archive for providing us with the ZIKV H/PF/2013 strain. We thank S. Lambert for helping to edit the paper. The work is supported by Singapore Ministry of Education Tier 3 grant (MOE2012-T3-1-008), National Research Foundation Investigatorship award (NRF-NRFI2016-01) awarded to S.-M.L. and the Duke-NUS Signature Research Programme funded by the Ministry of Health, Singapore.

Author Contributions E.X.Y.L. prepared the ZIKV sample for cryoEM work and carried out the plaque assays for testing the thermal stability of ZIKV and DENV. E.X.Y.L. and T.-S.N. optimized conditions for freezing samples on cryoEM grids. J.S. and T.-S.N. collected cryoEM images. G.F., S.Z. and J.S.G.O. boxed particles for cryoEM image reconstruction. T.-S.N. checked CTF curves of the micrographs. V.A.K. optimized and carried out the cryoEM image reconstruction procedure. G.F., S.Z. and V.A.K. fitted the structure into the cryoEM map. G.F., S.Z., V.A.K. and S.-M.L. analysed the structure. G.F., S.Z. and S.-M.L. prepared the figures. S.-M.L. and V.A.K. designed experiments. G.F., V.A.K., J.S., E.X.Y.L. and S.-M.L. wrote the manuscript. All authors discussed results and commented on the manuscript.

Author Information The cryoEM 3D reconstruction map of ZIKV has been deposited with the Electron Microscopy Data Bank (EMDB) under accession code EMD-8139. The modelled structure of the virus glycoprotein shell has been deposited with the Protein Data Bank (PDB) under the accession code 5IZ7. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.-M.L. (sheemei.lok@duke-nus.edu.sg).

Reviewer Information *Nature* thanks J. Huiskonen and Y. Modis for their contribution to the peer review of this work.

© 2016 Macmillan Publishers Limited. All rights reserved

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Virus plaque assay to determine the stability of ZIKV and DENV serotypes at different temperatures. The virus strains used for the plaque assay experiments were: DENV2 (NGC), DENV2 (PVP94/07), DENV4 (06K2270) and ZIKV (H/PF/2013). Virus samples were pre-incubated for 30 min at 4 °C, 29 °C, 37 °C and 40 °C and chilled at 4 °C for at least 30 min. Samples were serially diluted and added to BHK-21 (ATCC) cells grown in 24-well plates. For plaque assays, the cells were not specially authenticated, and not tested specifically for mycoplasma contamination. Virus was incubated with the cells for 1 h at 37 °C, then layered with methyl cellulose. After incubation at 37 °C for 4 days, the cells were fixed and stained to determine the plaque-forming units.

A similar experiment was done by pre-incubating ZIKV H/PF/2013 for 60 min at the same temperatures, before adding to BHK-21 cells.

Virus sample preparation. ZIKV strain H/PF/2013 was grown in *Aedes albopictus* C6/36 cells (ATCC) supplemented with RPMI1640 medium and 2% fetal bovine serum. The cells were inoculated with ZIKV at a multiplicity of infection of 2, and incubated at 29 °C for 3 days. Virus-containing medium was subjected to centrifugation at 12,227*g* for 45 min to remove cells and other debris, and the virus was precipitated with 8% (w/v) polyethylene glycol 8000 in NTE buffer (10 mM Tris-HCl, pH 8.0, 120 mM NaCl and 1 mM EDTA) overnight. Virus was pelleted by centrifugation at 14,334*g*, resuspended in NTE buffer and purified through a 24% sucrose cushion. Virus was further purified by ultracentrifugation using a 10–30% (w/v) potassium tartrate gradient. Virus band was extracted from the gradient with a syringe, and then the sample was buffer-exchanged into NTE buffer and concentrated using Amicon Ultra-4 100-kDa centrifugal concentrator (Millipore). Virus concentration and purity was estimated by Coomassie-blue-stained SDS–PAGE gel, where different concentrations of bovine serum albumin (BSA) solution were used as standards. All steps of the purification procedure were done at 4 °C.

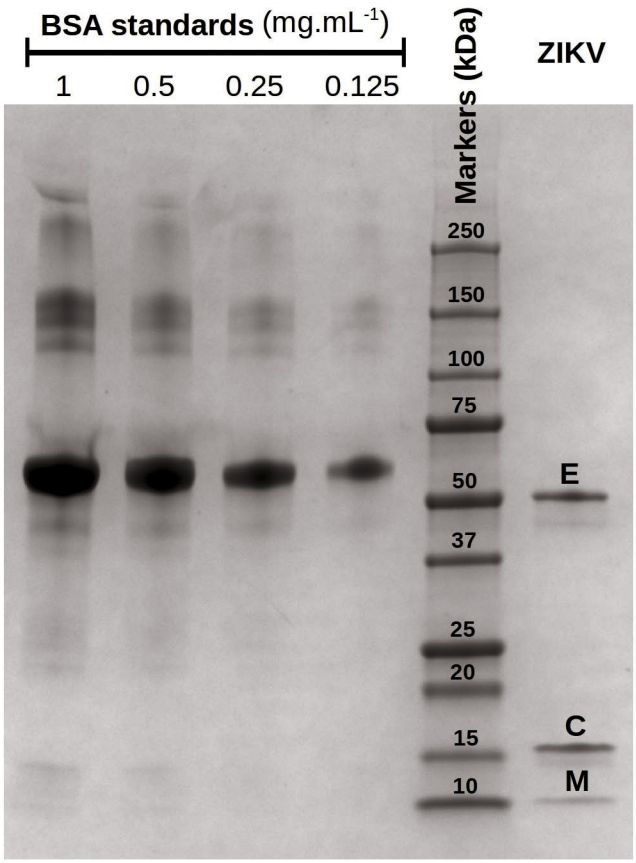
CryoEM sample preparation. Virus samples were incubated at 4 °C, 29 °C, 37 °C or 40 °C, for 30 min, then kept at 4 °C before freezing on cryoEM grids. About 2 μl of sample was put on a copper cryo-EM grid (with lacey carbon covered with thin carbon film), blotted for 1 s and plunged in liquid ethane bath using the Vitrobot plunger (FEI). The grids were stored in liquid nitrogen.

CryoEM data collection, image processing and 3D reconstruction. Images of vitrified virus were collected on a Titan Krios (FEI) microscope at 300 kV in nanoprobe mode using Leginon software for automated data collection. The magnification was 104,477, giving a pixel size of 1.34 Å. The images were recorded on Falcon II direct electron detector (FEI) in movie mode, with 30 frames per 1.8 s exposure time, giving a total dose of 50 e^{-} $Å^{-2}$. The individual frames were mutually aligned to correct for beam-induced sample drift with the MotionCorr program[26]. The first 10 frames were combined to produce the average image for further processing. The higher contrast full dose average was also calculated and used for defocus determination and for manual particle picking. In total, 11,600 individual virus particles from 3,083 micrographs were selected. The 3D reconstruction was carried out with MPSA[27], by using a 3D map of DENV4 (EMD-2485) as the starting model, with icosahedral symmetry averaging. The image data was split into two halves and processed with MPSA independently, to follow the 'gold standard' procedure for resolution determination. After 20 cycles of orientation search and 3D reconstruction, the final map was produced by using 7,180 particle images. The resolution of the map is 3.7 Å, at Fourier shell correlation cut off of 0.143 (Extended Data Fig. 2).

Model building and refinement. The resolution of the map allowed us to trace all of the polypeptide chains: three E and three M proteins. After manual structure building in Coot[28], using DENV2 coordinates (PDB 3J27) as a starting template, modified to include insertions, the protein structure had been refined with CNS[29] with icosahedral symmetry constraints applied, and showed R_{work} and R_{free} values of 37% and 39%, respectively. For refinement, the map has been 'sharpened' with application of a B-factor of −55 and low-pass filtered to 3.7 Å. Additional secondary structure restrains were applied to avoid overfitting. After refinement, the fit and the geometry of the protein structure was checked using Coot tools.

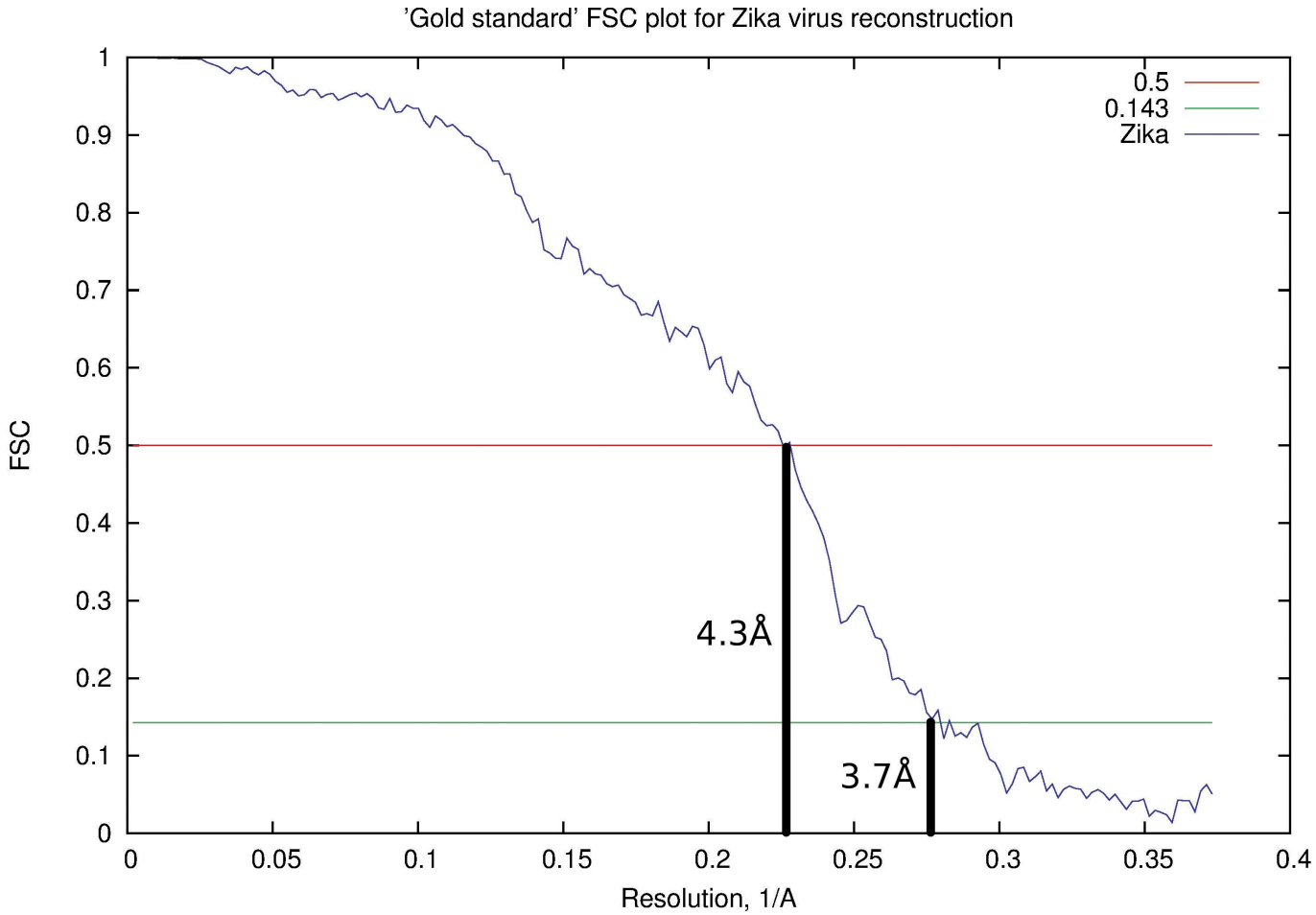
26. Li, X. *et al.* Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nature Methods* **10,** 584–590 (2013).
27. Liu, X., Jiang, W., Jakana, J. & Chiu, W. Averaging tens to hundreds of icosahedral particle images to resolve protein secondary structure elements using a Multi-Path Simulated Annealing optimization algorithm. *J. Struct. Biol.* **160,** 11–27 (2007).
28. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66,** 486–501 (2010).
29. Brünger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54,** 905–921 (1998).

© 2016 Macmillan Publishers Limited. All rights reserved



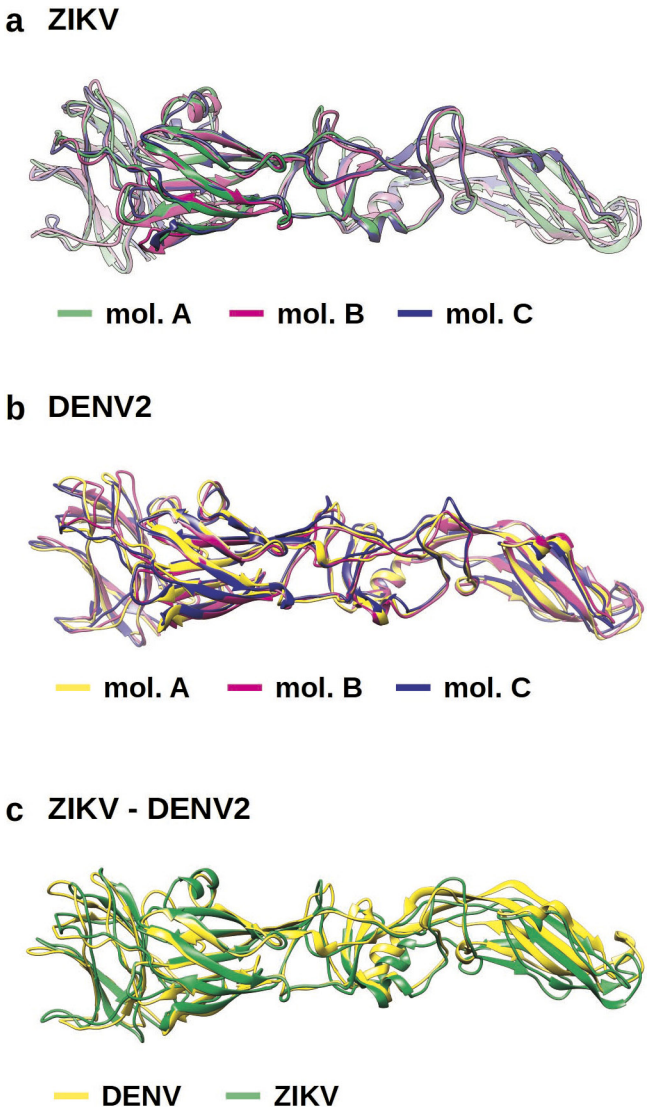
Extended Data Figure 1 | The Coomassie-blue-stained SDS–PAGE gel profile of ZIKV. The virus preparation showed clear protein bands corresponding to the E protein (50 kDa), capsid (C) protein (15 kDa), and M protein (10 kDa). The premembrane (prM) band (25 kDa) is not detected, indicating that there is little contamination by immature virus particles.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 2 | Gold standard Fourier shell correlation of the two cryoEM maps, made from odd and even numbered particles. At Fourier shell correlation (FSC) cut offs of 0.5 and 0.143, the resolution is 4.3 Å and 3.7 Å, respectively.

© 2016 Macmillan Publishers Limited. All rights reserved



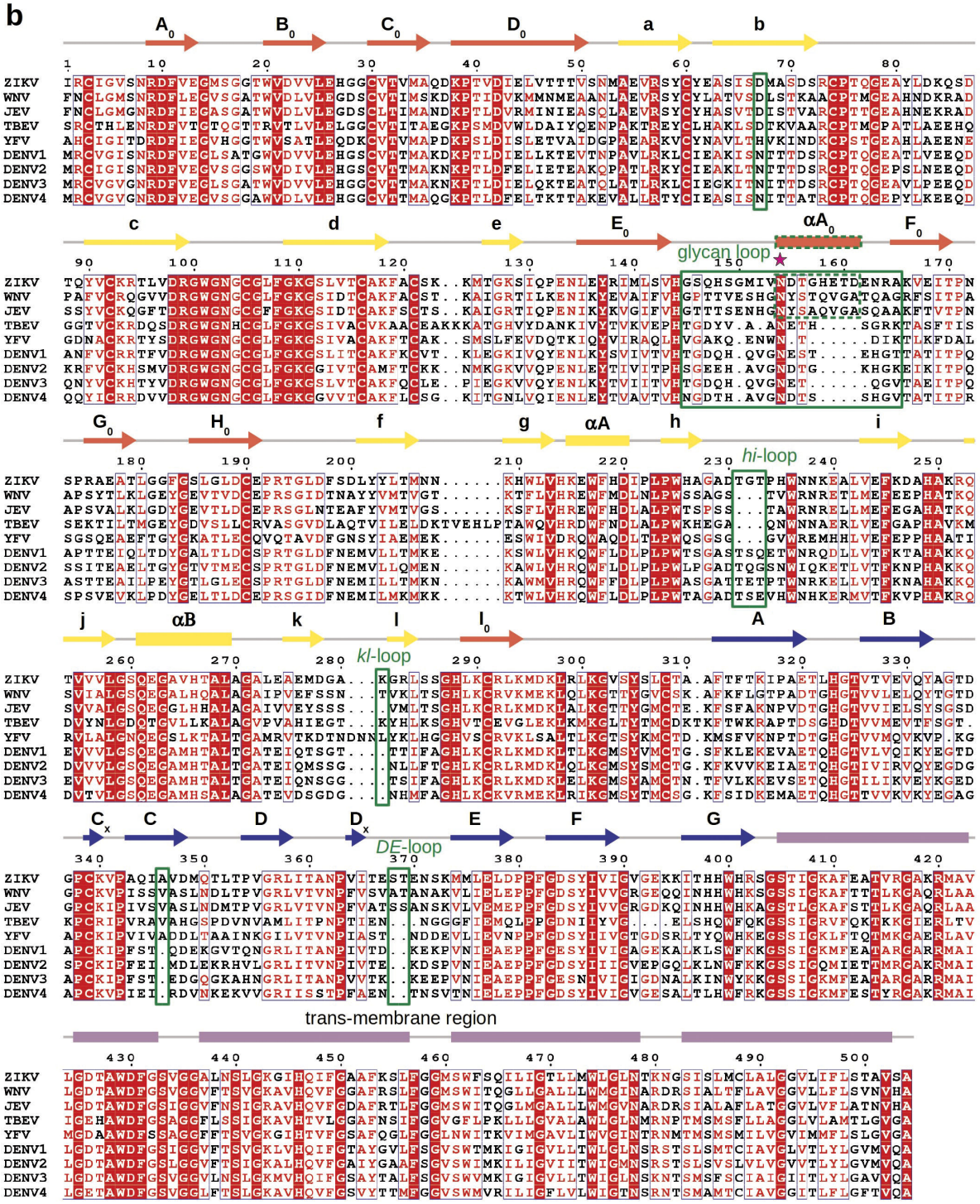
Extended Data Figure 3 | Comparison of ZIKV with DENV2 E proteins. **a**, Superposition of ZIKV E protein molecules A, B and C using DI as a reference point. The molecules are similar to each other. **b**, Superposition of the three individual DENV2 E proteins in an asymmetric unit by their DIs show high similarities between them. **c**, Superposition of molecule A of ZIKV with that of DENV2 by their DIs shows that they have slightly different DI–DIII and DI–DII hinge angles.

© 2016 Macmillan Publishers Limited. All rights reserved

a

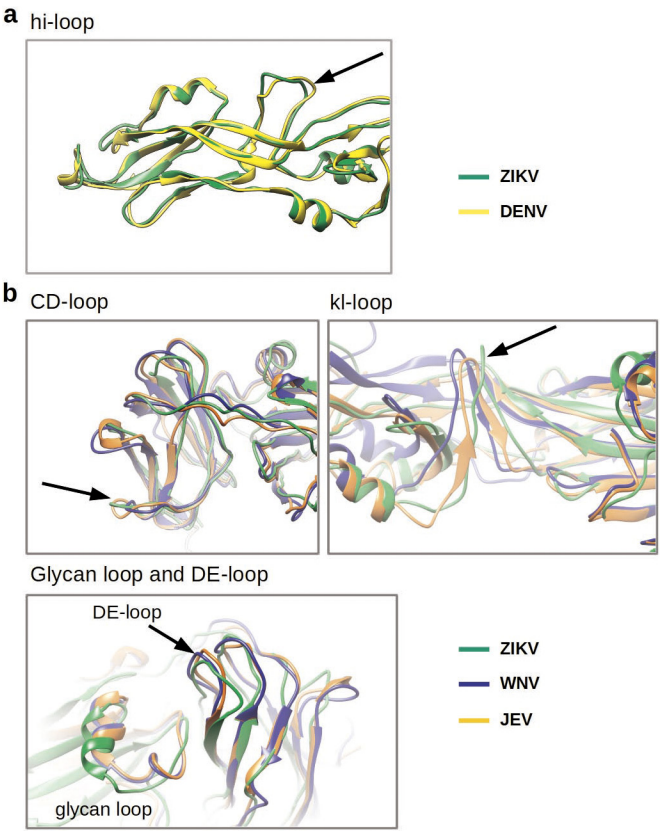
(%)	WNV	JEV	TBEV	YFV	DENV1	DENV2	DENV3	DENV4
ZIKV	54.09	54.20	40.25	42.45	57.37	54.34	58.01	56.57

b



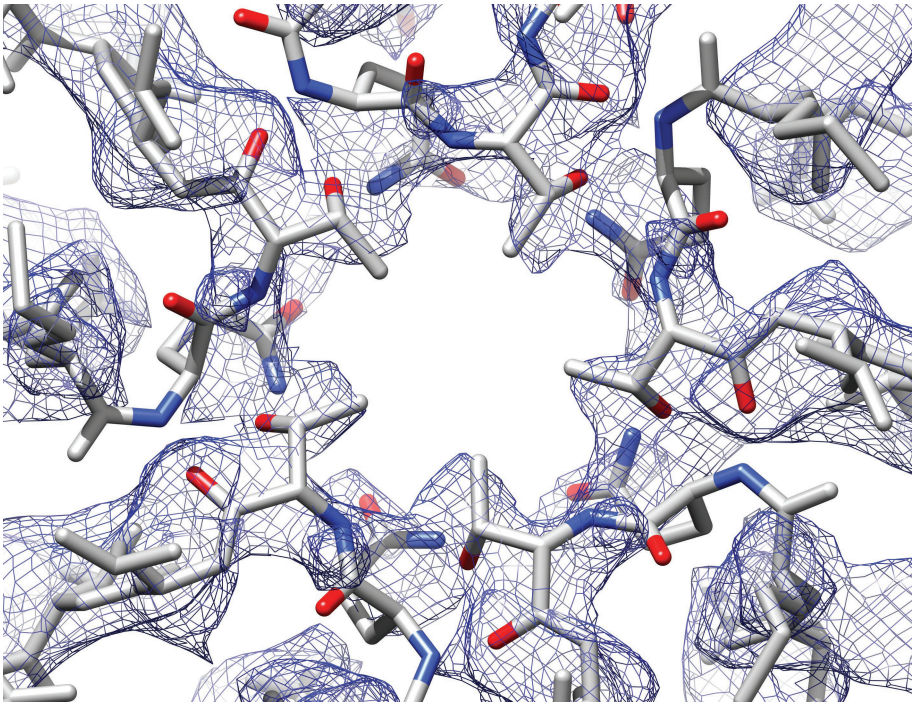
Extended Data Figure 4 | Sequence comparison of ZIKV to neurovirulent viruses (WNV, JEV and TBEV) and to febrile-illness-causing viruses (YFV and DENV1–4). a, **b**, Percentage sequence similarity (**a**) and sequence alignment (**b**) of flaviviruses. Green boxes indicate the most different regions between the neurovirulent viruses and the febrile-illness-causing viruses. Contributions of these regions to the virus pathology are unknown. ZIKV contains regions that are distinct in both virus groups.

© 2016 Macmillan Publishers Limited. All rights reserved



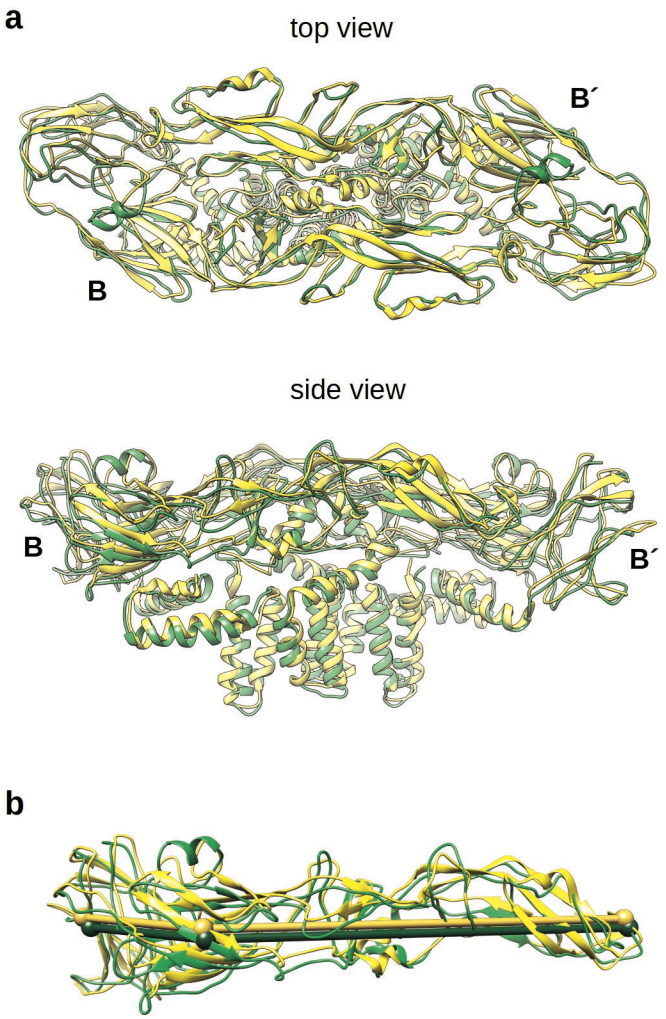
Extended Data Figure 5 | Structural comparison of different regions on the E protein sites unique to febrile-illness-causing virus and neurovirulant virus to ZIKV. a, Comparison of the hi-loop between ZIKV and DENV by superimposing their DIIs. **b**, Comparison of the glycan loop (DI), CD-loop (DIII) and DE-loop (DIII), between ZIKV, WNV and JEV by superimposing their DIIIs. Comparison of the kl-loop (DI–DII hinge) by superimposing these viruses by DI.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 6 | The fitted CD-loop in the cryoEM density around the five-fold vertex.

© 2016 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 7 | The ectodomain of the ZIKV B–B′ dimer is positioned lower in the radius compared to DENV2. a, Top panel shows the top view of the B–B′ dimer of ZIKV (green) and DENV2 (yellow), while the bottom panel shows the side view. The position of stem and transmembrane regions in both ZIKV and DENV are similar. **b**, The ZIKV E ectodomain is positioned at a slightly lower radius than DENV2. Stick connecting similar atoms in ZIKV and DENV is shown to help with the comparison of the positions of the E ectodomains.

© 2016 Macmillan Publishers Limited. All rights reserved

SCIENCE TO MARKET A neuroscientist takes on deal-making and tech transfer **p.434**

JOB SKILLS Teach yourself to love networking **go.nature.com/cwocoj**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

CAIAIMAGE/ROBERT DALY



Take time to make brief visits to faculty members, who can provide leads on grants, awards and jobs.

RELATIONSHIPS

Doctor's advice

US doctoral students may find that their dissertation committee can double up as a board of career mentors.

BY ALAINA LEVINE

PhD candidates in the United States have a certain advantage over students elsewhere. They are often required to assemble a dissertation committee — mentors who advise on topic selection and research ideas, scrutinize data, edit drafts and help to direct the work through submission and defence.

But members of a dissertation committee can often do a great deal more than that. They can lend emotional support or help to expand professional networks. They can offer leads and advice on grants, awards and jobs.

To maximize this relationship, the canny PhD student will select committee members with care and nurture the bond (see 'How to get the most out of your dissertation committee'). Those on both sides of the experience advise staying in close and regular contact with mentors from day one, then keeping them updated on your evolving interests and progress during the doctoral programme — and beyond. "If the committee is chosen wisely, these people can be pillars of support for years to come," says Jelena Kovačević, a biomedical engineer at Carnegie Mellon University in Pittsburgh, Pennsylvania.

Steven Senger talks of his dissertation-committee members in glowing terms. During his 2005–11 mathematics PhD, one found funding for him to attend several conferences and accompanied him to each, including a prestigious annual maths summit that unites researchers from academia, industry and government. Senger, who is now a mathematician at Missouri State University in Springfield, notes that his mentor "was careful to pick certain people whom he wanted me to meet and have lunch with". One such meeting spawned a collaboration with a prominent senior colleague.

Another of Senger's committee members explained how to cultivate collaborations and how best to conduct research in maths. "He walked me through the logical thought process," Senger says. "He didn't just regurgitate a paper to me — he explained how he got to this conclusion. He treated me like an apprentice. There's a difference between giving someone the bare facts and training them to be a peer."

HONESTY APPRECIATED

To glean the most from the relationship, it's best to be honest about what you want to do with your career, even if it differs from a mentor's own path — as maths PhD student Vyacheslav Kiria-Kaiserberg learned. In 2011, he asked mathematician Alex Iosevich of the University of Rochester in New York to be one of his committee members. The two subsequently spent many hours talking informally, and Kiria-Kaiserberg was upfront about his hopes to work in industry. So Iosevich introduced Kiria-Kaiserberg to his brother Sam, a manager at a data-science company in Chicago, Illinois, over dinner at Iosevich's home.

The informal meeting was a success by all accounts: this summer, after Kiria-Kaiserberg defends his dissertation, he will start work as a consultant at Sam Iosevich's company. "It's very important to choose the right people," says Kiria-Kaiserberg. "You can benefit from them and they can benefit from you."

Openness also helped medical researcher Luz Claudio in 1991 as she wrapped up her neuropathology PhD. She knew that she was more interested in applying her research to problems in the community than studying neurons at the bench, and had talked often of her interests with a committee member, a neuropathologist at her institution. He suggested a science-policy fellowship, and he wrote a recommendation that led to her winning one with the US Environmental Protection Agency (EPA).

After the stint ended, Claudio knew that she wanted to work at the intersection of public-health research and policy. Today, as head of the division of international health at Mount Sinai School of Medicine in New York City, Claudio has won renown for her work on asthma in ▶

© 2016 Macmillan Publishers Limited. All rights reserved.

TRADE TALK
Cell negotiator



Ben Thiede is a business-development officer at STEMCELL Technologies in Vancouver, Canada. As a neuroscience graduate student, he worked out methods to coax stem cells to take on specialized properties of the inner ear. Now, he works out intellectual-property and other deals that help to bring technologies to the scientific market.

When did you first think about jobs beyond the bench?
About two years before I graduated from my PhD programme at the University of Virginia in Charlottesville, I got married and started thinking more long term about my career. Before that, science was just fun.

What led you to your current career?
A classmate who was a year ahead of me was graduating with her PhD and going to law school. She had been exposed to other careers through an internship for the university's licensing and ventures group, working on technology transfer — helping to license and commercialize academic work. I did some informational interviews along the same lines, but I got an internship by talking to a guy who was fitting me for a suit. He was the father-in-law of one of the licensing managers at the university's patent foundation. I e-mailed the manager, and he took me on.

How did your internship compare to the lab?
As a grad student, you think narrowly about one problem. My learning curve in tech transfer was steeper than it was in the lab. I had to quickly pick up what was important about a technology: how does it fit into the field, and what is its value? It was really interesting for me to be close to science that was close to the point of being useful to people.

What do you do now?
I am looking for technologies, and evaluating whether they work or not and what is already patented. After we negotiate a license to bring in a technology, I can see that in six months or a year we are going to be using it in a product that researchers can buy and use in their labs. ■

INTERVIEW BY MONYA BAKER

This interview has been edited for length and clarity. See go.nature.com/xdttdg for more.

MENTOR MATCH
How to get the most out of your dissertation committee

- **Diversify** Populate your committee with people who are in your subfield and in related fields. Consider adding members who work at institutions other than your own, or outside academia.
- **Communicate early and often** Meet privately with your committee members so that you can develop a personal rapport. Check whether they prefer phone, text, e-mail or in-person contact.
- **Be honest and clear about your goals and concerns** Discuss your professional objectives, such as whether you want to work in academia or industry, move to a different city or nation or pursue a career in science, technology or elsewhere. The more they know about you, the more they can help you.
- **Stay in contact** Send e-mails, ask for coffee appointments, request short Skype chats and send holiday greetings.
- **Promote yourself** Keep committee members abreast of your progress, successes, challenges and aspirations so that they will keep you in mind for potentially game-changing career opportunities.
- **Pay it back** Look for ways to help them, too. You have a network and access to information, ideas and opportunities. Work out how you can assist them with their career and professional interests. A.L.

▶ low-income communities. She credits her relationship with the committee member: "That experience changed my life."

Although frequent face time is important, PhD students should choose at least one committee member who is outside their own institution, says environmental scientist Tammy Newcomer Johnson. That external connection, she says, can help to maximize potential career gains. Her own adviser at the University of Maryland in College Park encouraged her to diversify her committee, and she invited two 'outsiders', including Paul Mayer, an EPA ecologist whom she had met during fieldwork. They discussed career possibilities in the federal government, and Mayer noted links between Newcomer Johnson's doctoral research and the EPA headquarters in Washington DC — where she is now a postdoc.

Thanks to a recommendation from him, she was invited to present a seminar at an EPA lab on monitoring stream restoration. Her talk boosted her CV and led to new contacts in the agency, who in turn gave her more information about research posts there. Now she is mulling over staying with the agency for the long haul. "It gave me a sense of what a scientist does at the EPA — I realized that this could be a really wonderful career path," she says.

Contact with committee members can stretch far beyond those PhD years, as can the benefits. Kovačević has sat on about 30 dissertation committees; one of her former students turned to her for advice half a dozen years later while pursuing an assistant-dean position at a major US university. Kovačević proofread the CV and suggested improvements. She has also nominated another former student for membership of a prestigious electrical-engineering committee.

Indeed, there is much for a PhD student to gain from a warm association with committee members, but the best former students will also consider what they can give back. In January, Newcomer Johnson invited members of her former committee to give talks at the EPA. "I'm connecting them with my network," she says. A former student of Kovačević's landed a desirable position at a company and has since made sure to ask the professor to circulate the company's job openings to her students. "This is a two-way benefit," says Kovačević. "She gets help to succeed at her job, and the benefit to me and my institution is placing students at a good company."

Students should remind faculty members whom they invite to serve on their dissertation committee that they are likely to benefit. "One measure of success for professors is the success of their students," says Claudio. "When you are up for promotion and tenure, the better your students do, the better you look."

In the end, seasoned committee members say, the rewards of a dissertation committee can go both ways. "Understand this is an incredible potential resource," says Kovačević. "Do as much as you can to nurture it." ■

Alaina Levine *is a freelance science writer based in Tucson, Arizona.*

CORRECTIONS

The Careers Feature 'Take my advice' (*Nature* **532,** 531–533; 2016) wrongly implied that the Houston-based consulting club is run by the Texas Medical Center. In fact, it is an independent organization.
The Careers Feature 'Change is in the air' (*Nature* **532,** 403–404; 2016) wrongly affiliated climate-change ecologist George Divoky with the University of Alaska Fairbanks. He is actually now director of Friends of Cooper Island in Alaska.

© 2016 Macmillan Publishers Limited. All rights reserved.

PROJECT EARTH IS LEAVING BETA

Get ready for a brand new experience!

BY J.W. ALDEN



ILLUSTRATION BY JACEY

Dear Backers,

The beta test of *Project Earth* is finally over. We can't thank you enough for your patience and generosity during our crowdfunding period. But before we roll out *Earth 1.0*, there's a few things you should know.

First, the bad news: this update comes with a server wipe. Yes, we know you've poured time and effort into your 'lives' on Earth, and it's disappointing to lose your progress. Unfortunately, this can't be helped. We experimented with methods of porting existing avatars into the new version, but it brings a host of compatibility issues with our new character-creation system (more on that later). As a consolation, we're planning another exciting in-game 'End of the World' event, so you can go out with a bang.

Here are some big changes to expect in the new update:

NEW: Abuser Report System Every multiplayer experience comes with griefers and trolls. Unfortunately, the nature of Earth's gameplay makes it very difficult to tell when someone is being an asshole *on purpose*, but we've implemented a system that lets you flag an avatar for review. This notifies our mod team, who will examine the player's behaviour for signs of intentional griefing. Please don't abuse this feature! Only report an avatar if it exhibits unreasonable hostility that could not possibly be the result of normal 'human' development.

NEW: Advanced Career Paths A limited number of political and business leadership careers will open to the general playerbase. Previous builds allowed only members with mod privilege to take on certain leadership positions in-game, which allowed us to shape the progress of 'civilization' as we saw fit. We've listened to your feedback on this policy, and we're thrilled to open this aspect of the game to all players. We're also hoping this will address the 'conspiracy theory' phenomenon that became prevalent in previous builds.

NEW: Space Exploration Many of you were quite vocal when we curtailed the space race during our 'cold war' in-game event. At the time, we worried about the playerbase advancing past the capabilities of our current build, so we steered focus back towards Earth itself. Since then, we've been working hard on a new procedural engine for the cosmos, and now we're ready to roll it out. We hope this will revitalize many career questlines, including the astronaut path. Oh, and don't think we haven't heard your complaints about the light-speed barrier. These things take time.

FIXED: Character Imbalance We turned a blind eye to this for a while, because we knew a fix would cause waves, no matter how we approached it. But after long deliberation, we've given the character-creation suite an overhaul. A proprietary algorithm now determines your character's congenital traits, including things such as gender, race and sexual orientation. Any trait your avatar is 'born' with will be free of manipulation from players or mods. We know this might upset those of you who feel these traits are a matter of player choice, but we believe this is the more realistic approach. An increase in diversity will only enhance our experience on Earth. Plus, the inexplicable dominance of the 'straight white male' template kept crashing parts of the system.

FIXED: Dreams We've finally isolated the cause of the 'dream' phenomenon, in which an avatar retains glimpses of the player's real-world memories upon waking from logout status. Some team members argued that we should make this bug a supported feature, owing to the impact of dreams on art and culture in *Earth 0.91*. In the end, we're not comfortable with the risk involved. Although it hasn't happened yet, this dream glitch *could* lead to avatars realizing that their 'lives' are a game. We're not entirely sure what would happen if this occurred, but we're looking to avoid the sudden emergence of a secret resistance group working towards exposing the simulation, so we decided to remove them.

FIXED: History Leak There's been discussion on the forums about a 'history leak' phenomenon, in which various myths, legends and religious tales seem to reflect previous builds of *Earth*. For instance, the Greek pantheon resembles the superuser avatars that the development team used in our alpha test. Many flood myths bear a strong resemblance to the 'End of the World' event from our first playable build. While working on the dream bug, we discovered a crucial link between these two phenomena, which helped us to iron this one out. Sorry, but your old avatar won't be worshipped as a messiah in *Earth 1.0*.

As this marks our official release, we anticipate a large influx of new players. We've set up a special section of the forums for first-timers. Please consider paying it a visit and offering some advice to the noobs. Remember how disorienting your first trip to Earth was (especially if you installed the 'childhood' DLC expansion).

The new 'End of the World' event is scheduled for 21 December. We expect a doomsday cult or two to arise before then, as some of you will unwittingly communicate this date to your avatars in dreams. For this reason, we're keeping the nature of the apocalypse a secret until the end is upon you. Rest assured we've come up with something exciting and different this time around. No floods, plagues or supervolcanoes. We'll give you one hint: watch the sky.

On behalf of everyone here at Blue Marble, we'd like to thank our backers and beta testers. Our little world couldn't have made it this far without you, and we hope that you're all permanent residents. *Earth 1.0* is only the beginning!

E. D. Amun

Development lead, *Project Earth*

Blue Marble Games, Subsidiary of Ogdoad Entertainment. ■

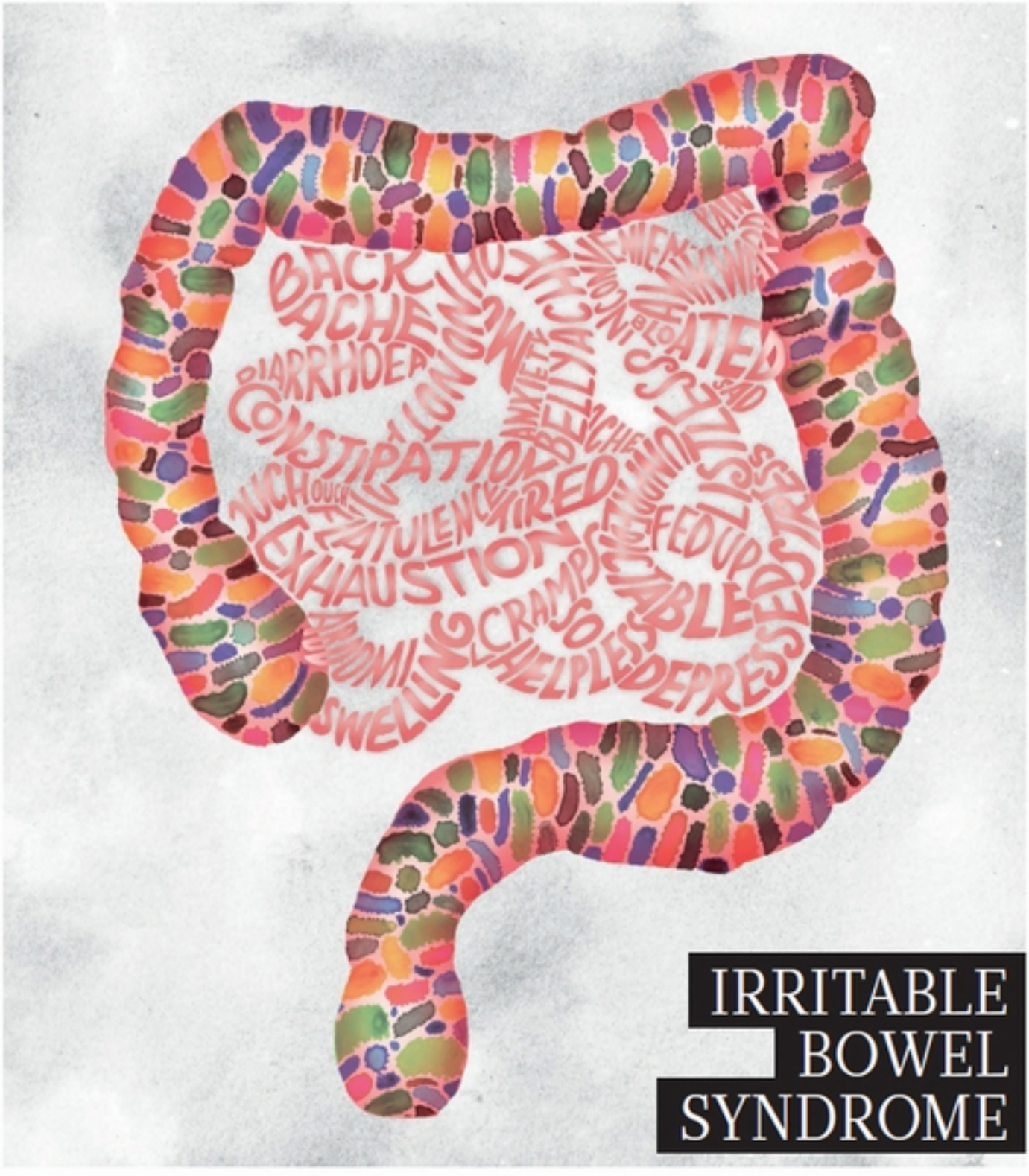
J. W. Alden *is a graduate of Odyssey Writing Workshop and a 1st Place Writers of the Future winner. He lives in Florida with his wife, Allison. Find him at www.AuthorAlden.com.*

NATURE.COM
Follow Futures:
@NatureFutures
go.nature.com/mtoodm

© 2016 Macmillan Publishers Limited. All rights reserved.

nature OUTLOOK

19 May 2016
Supplement to Nature
Publishing Group journals



IRRITABLE BOWEL SYNDROME

Untangling a complex condition

Produced with support from:

Allergan

© 2016 Macmillan Publishers Limited. All rights reserved.

nature OUTLOOK

IRRITABLE BOWEL SYNDROME

19 May 2016 / Vol 533 / Issue No 7603



Cover art: Sarah J. Coleman

Editorial
Herb Brody
Michelle Grayson
Richard Hodson
Jenny Rooke

Art & Design
Wesley Fernandes
Mohamed Ashour
Annthea Lewis

Production
Karl Smart
Ian Pope
Matthew Carey

Sponsorship
David Bagshaw
Yvette Smith

Marketing
Nicole Jackson

Project Manager
Anastasia Panoutsou

Art Director
Kelly Buckheit Krause

Publisher
Richard Hughes

Editorial director, partnerships media
Stephen Pincock

Chief Magazine Editor
Rosie Mestel

Editor-in-Chief
Philip Campbell

Nobody likes to talk about their bowel movements. This squeamishness means that irritable bowel syndrome (IBS), which entails a change in stool form and frequency, is widely under-reported. The condition is comprised of several symptoms, such as diarrhoea or bloating, that are also present in other diseases, complicating efforts to study and treat it. IBS can also be associated with mental-health conditions such as anxiety or depression. In the past, this led some clinicians to categorize it as mainly a psychiatric concern, which for many years also attracted unwarranted stigma. This Outlook describes these difficult issues, the progress made and the challenges that remain (see page S118).

IBS as we now know it has a short history, but it is heavily influenced by work that goes back centuries (S102). Diagnostic capability took a leap forward in the 1980s with the Rome criteria — the fourth iteration is published this month (S107). But what many people want is a reliable diagnostic test (S110), which would also lead to better, more targeted drugs (S116). One subtype of IBS caused by gastroenteritis might already be treatable — or even preventable (S114).

Food can trigger symptoms, and a diet low in FODMAPs — undigestible carbohydrates — has gained many supporters. But not all researchers are convinced of its effectiveness (S108). Something does happen in the large intestine, however, and researchers are trying to decipher the influence of the microbiome on the gut–brain axis (S104). Whatever the cause, one thing is clear — although IBS is mild for many people, for 1 in 20 patients it is a severe and even life-threatening condition. But help is available, and one clinician says that the first step is simply to listen to patients (S112).

We are pleased to acknowledge the support of Allergan plc in producing this Outlook. As always, *Nature* retains sole responsibility for all editorial content.

Michelle Grayson
Senior supplements editor

Nature Outlooks are sponsored supplements that aim to stimulate interest and debate around a subject of interest to the sponsor, while satisfying the editorial values of *Nature* and our readers' expectations. The boundaries of sponsor involvement are clearly delineated in the Nature Outlook Editorial guidelines available at go.nature.com/e4dwzw

CITING THE OUTLOOK
Cite as a supplement to *Nature*, for example, *Nature* **Vol. XXX**, No. XXXX Suppl., Sxx–Sxx (2016).

VISIT THE OUTLOOK ONLINE
The *Nature Outlook Irritable Bowel Syndrome* supplement can be found at http://www.nature.com/nature/outlook/ibs
It features all newly commissioned content as well as a selection of relevant previously published material.

All featured articles will be freely available for 6 months.

SUBSCRIPTIONS AND CUSTOMER SERVICES
Site licences (www.nature.com/libraries/site_licences): Americas, institutions@natureny.com; Asia-Pacific, http://nature.asia/jp-contact; Australia/New Zealand, nature@macmillan.com.au; Europe/ROW, institutions@nature.com; India, npgindia@nature.com. Personal subscriptions: UK/Europe/ROW, subscriptions@nature.com; USA/Canada/Latin America, subscriptions@us.nature.com; Japan, http://nature.asia/jp-contact; China, http://nature.asia/china-subscribe; Korea, www.natureasia.com/ko-kr/subscribe.

CUSTOMER SERVICES
Feedback@nature.com
Copyright © 2016 Nature Publishing Group

CONTENTS

S102 HISTORY
Transit time
Physicians have been treating gut problems for centuries

S104 MICROBIOME
Bacterial broadband
Microbes' influence on the brain

S107 PERSPECTIVE
An easier diagnosis
Brian E. Lacy discusses the Rome criteria

S108 DIET
Food for thought
The rise in popularity of the low-FODMAP diet

S110 DIAGNOSTICS
Filling in the missing pieces
Scientists search for IBS biomarkers

S112 Q&A
A listening ear
Peter Whorwell explains his approach to IBS treatment

S114 INFECTIOUS DISEASE
Something in the water
Pathogens could alter the microbiota

S116 DRUG DEVELOPMENT
A healthy pipeline
The IBS drugs on their way to the clinic

S118 RESEARCH
4 big questions
The challenges still facing scientists

COLLECTION

S120 **'Functional' gastrointestinal disorders—a paradigm shift**
Nicholas J. Talley

S122 **Primeview: Irritable bowel syndrome**

S124 **Intestinal microbiota and diet in IBS: causes, consequences, or epiphenomena?**
Mirjana Rajilić-Stojanović et al.

S134 **Altered colonic bacterial fermentation as a potential pathophysiological factor in irritable bowel syndrome**
Tamar Ringel-Kulka et al.

S142 **Neuronal correlates of placebo in chronic FGIDs**
QiQi Zhou & G. Nicholas Verne

© 2016 Macmillan Publishers Limited. All rights reserved.



Transit time

MIND AND BODY

460–322 BC

Medical science can be traced back to Hippocrates in ancient Greece. He and his contemporaries — Plato and Aristotle — develop the concept of holism: the mind and body are one, and medicine should treat both. For example, Hippocrates not only recommends physical treatments such as enemas (**pictured**), but also soul-enriching activities such as listening to music. Dysentery, which is now recognized as a trigger of IBS, was then thought to result from an overabundance of blood.



THOTH'S BEAK 3500–2500 BC

The ancient Egyptian ibis-headed god Thoth (**pictured**) was associated with the enema — ibises were thought to self-administer using their long beaks. The Egyptians had specialized physicians to treat abdominal problems.

1637

French philosopher René Descartes proposes the theory of dualism: the spirit and body are distinct entities. Because the spirit is no longer thought to inhabit the body, autopsies become more acceptable. Physicians uncover the pathological roots of many conditions and begin to develop treatments. But superstition is not banished altogether. Conditions with no identifiable pathology, including mental-health disorders such as schizophrenia, are blamed on possession and, later, insanity.

BIOFEEDBACK 1884

American psychologist William James proposes a direct link between emotions and the body. For example, the theory suggests that the feeling perceived as anxiety results from the sensation of butterflies in the stomach.

1800s

Several physicians publish accounts of bowel complaints with names such as membranous enteritis and colitis. "The bowels are at one time constipated, at another lax, in the same person. How the disease has two such different symptoms I do not profess to explain," writes William Cumming in the *London Medical Gazette,* adding, "The mental phenomena consequent on it amount, in some cases, to absolute insanity." Treatments included dietary changes, opiates, mustard poultices and electrotherapy.

ABOMINABLE ABDOMINALS 1920s

Some physicians apply derogatory terms to people who repeatedly seek help for digestive symptoms. In his lectures on dyspepsia, British physician Robert Hutchison complains of the "incessant" demands of the "abdominal woman".



1945

The Second World War is over, but not the intestinal troubles of many soldiers and sailors who contracted dysentery in the tropics. Their symptoms continue or recur even after the causative microbes have gone. Writing in the *British Medical Journal*, G. T. Stewart suggests that long-term changes to intestinal flora might contribute to lingering diarrhoea.

MOULD BREAKING 1928

British biologist Alexander Fleming discovers a mould that kills bacteria, leading to the development of penicillin and other antimicrobials. These drugs will help to tame the cause of dysentery.

ROAD TO ROME

1962

Nazir Chaudhary and Sidney Truelove describe what they refer to as irritable colon syndrome. To their patients, the pair call the problem misbehaviour of the bowel and treat it with the sedative phenobarbital and the muscle relaxant propantheline bromide. Patients also report psychological symptoms such as anxiety or depression. Over the decade, the term irritable bowel syndrome becomes more popular.

PRESIDENTIAL PROBLEMS 1961

During his time in office, then US president John F. Kennedy suffers from severe diarrhoea — as he did for much of his life. Doctors diagnose spastic colitis, which would now be called IBS.

CLOCKWISE FROM TOP LEFT: JOSEPH KUHN REGNIER/ADOC-PHOTOS/CORBIS; ROGER WOOD/CORBIS; WORLD HISTORY ARCHIVE/TOPFOTO

© 2016 Macmillan Publishers Limited. All rights reserved.

The term irritable bowel syndrome (IBS) is relatively modern, but people have had similar symptoms for millennia. Initially thought to exist only in the mind, the disorder has gained legitimacy through the identification of causes and improved diagnosis. By **Amber Dance.**

RELAX 1965

The first drug marketed specifically for IBS in Europe is the anti-spasmodic mebeverine, which relaxes the gut muscles.



TUMMY BUG 1984

Australian physician Barry Marshall drinks *Helicobacter pylori* to prove that the bacterium causes gastritis and ulcers, disproving the long-held view that these conditions have a purely psychological origin.

STRENGTH IN NUMBERS 1987

The patient organization the IBS Self Help and Support Group is formed in the United States. It now has around 60,000 members.

DUALISM DEBUNKED 1988

The idea that IBS can result from interactions between the brain and the gut emerges from a survey reporting that certain psychological conditions typically make the syndrome worse, and lead people to seek medical attention (D. Drossman *et al. Gastroenterology* **95,** 701–708; 1988).

HEART HURTS 2007

Pharmaceutical company Novartis withdraws tegaserod, a drug for constipation-predominant IBS, after reports that it can cause severe cardiac problems in some people.

IN THE GENES 2014

About 2% of people with IBS carry mutations that cripple the *SCN5A* gene, which is known to encode a sodium channel that acts as a pacemaker for contractions in both the gastrointestinal tract and the heart (A. Beyder *et al. Gastroenterology* **146,** 1659–1668; 2014).

L: PHOTO ITAR-TASS/ YURY BELINSKY/ALAMY; R: JASON MERRITT/GETTY

1987

Gastroenterologists gather in Rome to draft guidelines for the diagnosis of IBS, helping to legitimize the condition. The criteria generate some controversy: they do not perfectly distinguish IBS from other similar conditions, and they are difficult to validate. But use of these guidelines grows in clinics, research and drug trials, and grant agencies insist on their use (see page S107).

2000

In February, the US Food and Drug Administration (FDA) approves the serotonin-receptor blocker alosetron for women with diarrhoea-predominant IBS. By November it has been withdrawn. Several patients have had complications due to constipation, such as a ruptured bowel, and some have died. Under public pressure, the FDA allows its use again in 2002, but only for women with severe diarrhoea.

2004

Researchers describe the first evidence of a specific molecular defect that may be a cause of IBS. The guts of people with IBS, as well as those with ulcerative colitis, contain relatively low amounts of both serotonin and an enzyme that synthesizes it, as well as a transporter that brings serotonin into cells (M. D. Coates *et al. Gastroenterology* **126,** 1657–1664; 2004). Although this result is initially controversial, later studies confirm it.

BACTERIAL BOUNTY

2008

Evidence emerges that the low-FODMAP diet, which omits foods containing short-chain carbohydrates, can improve symptoms in people with IBS (S. J. Shepherd *et al. Clin. Gastroenterol. Hepatol.* **6,** 765–771; 2008). It is a very restrictive diet, but, because it seems to help up to two-thirds of people, it is increasingly adopted, often without medical advice (see page S108).



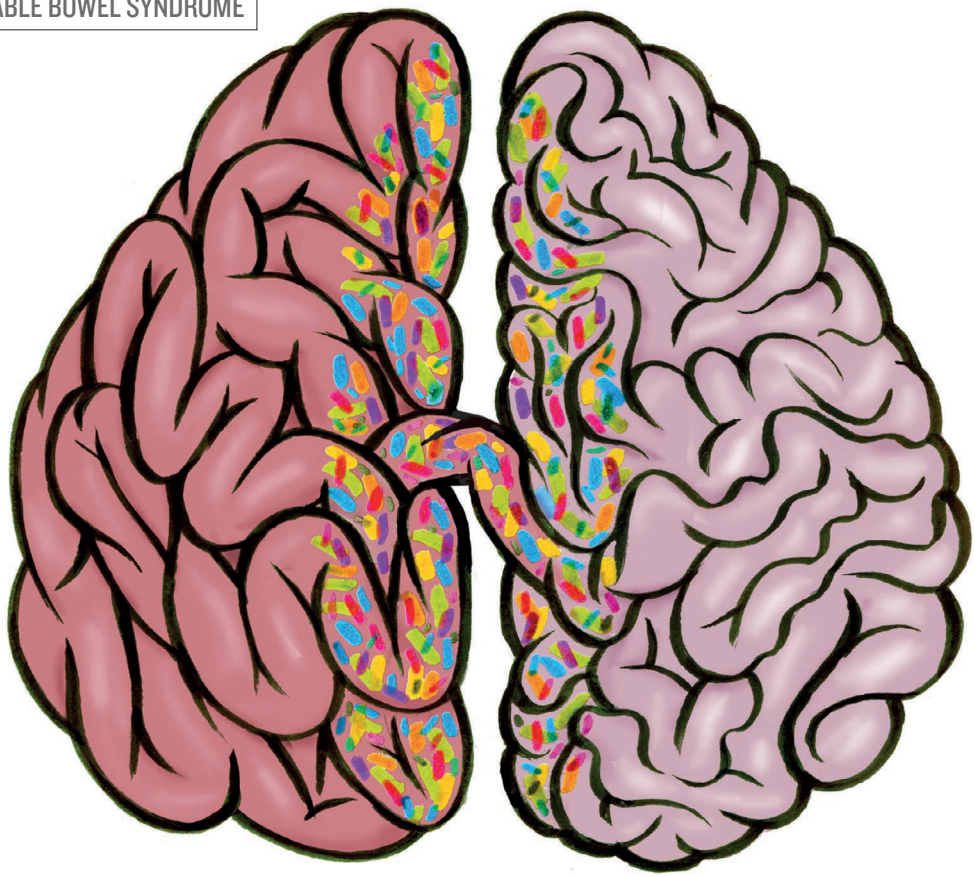
2011

Faecal analysis reveals that people with IBS have an altered gut microbiome, including more representatives of the phylum Firmicutes and fewer of Bacteroidetes (see page S104). This discovery opens up possibilities for diagnostic tests and treatments. Later, children with IBS who have certain microbes in their gut are found to be more likely to benefit from the low-FODMAP diet.

2016

The stigma surrounding IBS continues to fade. During the Super Bowl, football fans are shown an advertisement for the drug rifaximin, featuring an animated intestine. Celebrities such as supermodel Tyra Banks (**pictured**) have previously spoken about their digestive symptoms. But despite the recent approval of several medications, scientists have much more to do to untangle the disparate causes of IBS.

© 2016 Macmillan Publishers Limited. All rights reserved.



MICROBIOME

Bacterial broadband

The involvement of intestinal bacteria in gut–brain communication could help to explain the mysteries of irritable bowel syndrome, but the search continues for definitive evidence.

BY MICHAEL EISENSTEIN

Most people prefer to imagine a healthy distance between their brain and colon, but in fact the two are intimately linked. Neuronal lines of communication help the brain to control the digestive process, and messages relayed back by the gut can influence both perception and behaviour.

Evidence for another important player in this dialogue is coming to light. The human digestive tract is host to a massive and diverse community of bacteria that aid digestion and strengthen immunity — and over the past decade, scientists have observed that the gut microbiota also interact with the central nervous system (CNS). As a result, mental health and even neurological development might both shape and be shaped by the composition and behaviour of these bacteria.

Symptoms as diverse as gastrointestinal distress, physical pain and depression may all converge on the health of our bacterial communities. This could be especially relevant in irritable bowel syndrome (IBS), in which physical effects such as constipation and diarrhoea often co-exist with psychiatric problems such as anxiety and post-traumatic stress. Animal experiments offer some evidence of bacterial involvement in the condition, whereas studies in people with IBS show that a dose of healthy probiotic bacteria could mitigate both digestive discomfort and psychiatric symptoms.

These and other findings suggest that the microbiota act as crucial 'first responders', relaying health-related signals between the digestive tract and the brain. Perturbations in this gut–brain–microbiome axis may play a central part in IBS, but much of the evidence for this is patchy or circumstantial. Many questions remain about the extent to which this axis influences IBS pathology, about whether defects in microbiota are a cause or a consequence of the condition, and about how to turn this knowledge into treatments.

MIND–BODY CONNECTION

The existence of a brain–gut axis is not a particularly avant-garde idea. The roots of the concept go back to the late nineteenth century, when the physicians William James and Carl Lange first proposed that emotional response might be directly modulated by signals transmitted from the viscera to the brain (see page S102).

The intestinal wall maintains bidirectional communication with the CNS through the vagus nerve, which runs from the brainstem to the gut and manages numerous unconscious tasks throughout the body, such as maintaining heart rate. The vagus nerve also relays essential stress-response signals. Its interaction with the gut regulates almost every aspect of the passage of digested material through the intestines — explaining why CNS disorders are so often accompanied by digestive problems.

"If you have alterations in the brain, you will almost certainly have altered output to the gut because the two organs are that closely connected," says Emeran Mayer, a gastroenterologist at the David Geffen School of Medicine, University of California, Los Angeles. Conversely, feedback signals from the gut to the brain can go well beyond obvious hungry-versus-full messages; abnormalities in the digestive system can directly shape both cognitive and emotional state.

Research over the past few decades has clarified the role of this gut–brain axis in IBS. For example, brain-imaging studies by Mayer and colleagues have identified both structural and functional alterations in the CNS associated with IBS, including notable changes in specific brain regions that respond to pain and

SARAH J. COLEMAN

© 2016 Macmillan Publishers Limited. All rights reserved.

discomfort[1]. Mayer's group has also identified top-down effects of psychosocial trauma and post-traumatic stress — especially from early in life. "In patients with IBS," says Mayer, "up to 60% indicate that stress is the factor that triggered either the first onset or the exacerbation of symptoms."

Furthermore, many people with IBS have associated psychiatric issues, and some find that mental-health-oriented treatments can improve their gastrointestinal symptoms. "Cognitive behavioural therapy, hypnotherapy and psychodynamic therapy are all quite effective," says Magnus Simrén, a clinical gastroenterologist at the University of Gothenburg in Sweden. And although trials of such interventions for IBS have proved challenging to perform and interpret, he notes that "they have still been quite impressive when it comes to their effectiveness in treating IBS".

Numerous lines of evidence also link IBS to perturbations of the gut microbiome. Many IBS cases arise after a gastrointestinal infection, which can markedly change the composition of the gut microbiota both directly, by killing off certain beneficial species, and indirectly, by damaging the intestinal environment that supports them (see page S114). Drugs can also influence the health of the microbiome. John Cryan and his colleagues at University College Cork in Ireland have shown that rats treated early in life with a cocktail of antibiotics that kills off certain types of gut microbe can experience long-lasting gut hypersensitivity[2] — a common symptom of IBS.

However, the IBS microbiome data are still swimming into focus. Each gram of human colon tissue can contain up to one trillion bacteria, representing at least ten major categories (or phyla) that comprise many poorly characterized subtypes. This means that most microbiotal data represent high-level surveys rather than detailed censuses.

Microbiologist Willem de Vos of Wageningen University in the Netherlands was among the first to profile microbiome changes in large numbers of people with IBS[3]. His team's surveys revealed some notable trends, including a reduction in the numbers of 'good' species, such as those that belong to *Bifidobacterium*, that typically form the bedrock of a healthy microbiota, and an increase in *Ruminococcus* species that are associated with infection. "We've seen some patterns, and there's reason to believe that there are robust signatures of IBS," he says. But the results so far are conflicting. Some researchers have observed reduced microbial diversity in IBS, whereas others — including Mayer's group — have seen the opposite.

Healthy microbiota are thought to collaborate with host epithelial cells to maintain the intestinal wall. Destabilization of this relationship could compromise the gut and impair digestion or excretion — defining symptoms of IBS. De Vos, together with Robin Spiller at the University of Nottingham, UK, showed that alterations to the microbiota in people who developed IBS after an infection were accompanied by striking changes in intestinal gene expression[4]. Importantly, these correlated with alterations in specific subpopulations of gut bacteria. Moreover, these changes seem to be self-perpetuating in people with IBS.

"Some people with *Campylobacter* infections get diarrhoea for a while and then go back to normal," de Vos says. "But others don't go back and have recurrent problems." In other words, he says, their microbiome falls into what is known as an "alternative stable state" that locks them into IBS.

A BACTERIAL BRIDGE

Only in the past few years have researchers come to suspect that all three elements — gut, microbiota and brain — intersect in IBS. Early evidence for a brain–microbiome connection came in 2011 from scientists at the Karolinska Institute in Stockholm. They demonstrated that 'germ-free' rodents that have been reared under sterile conditions and lack a microbiota have a reduced anxiety response in various laboratory tests relative to those with intact bacterial communities[5]. Other studies with germ-free animals have shown that disruptions in the microbiota can increase sensitivity to pain and bring on cognitive symptoms that resemble depression and anxiety.

New evidence sheds light on the mechanism of the gut microbiota's impact on nervous-system function. Researchers in Japan have found that germ-free mice produce lower than normal levels of dopamine and adrenaline[6], two important molecules in the CNS response to stress. The microbiota also influence production of serotonin — an intriguing finding given that this neurotransmitter plays a central part in gut motor function and digestion, as well as in various cognitive and mood disorders.

"The animal data are almost hard to believe because they're so dramatic."

Roughly 95% of the body's serotonin is produced in the digestive tract, and a team led by Elaine Hsiao, now at the David Geffen School of Medicine, has demonstrated in mice that a particular subset of gut bacteria directly stimulates intestinal serotonin synthesis and release[7]. "This is one of the most interesting papers in this field from the last ten years," says Mayer. "Those serotonin-producing cells in the gut can be viewed as the main hub for communication between the gut and brain, because they're exposed to what we eat and to signals produced by the gut microbes, and innervated by connections to the vagus nerve."

Several mouse studies suggest that the vagus nerve serves as some sort of 'hotline' by which gut microbes communicate directly with the CNS (see 'A complicated conversation'). Cryan and his colleagues have found that the effects on the CNS of a probiotic bacterial strain seem to depend on signals relayed by the vagus nerve[8]. "That immediately tells us that vagal pathways may be relevant to IBS, and the microbiome could be shepherding that," Cryan says. A team led by Premysl Bercik, a gastroenterologist at McMaster University in Hamilton, Ontario, has likewise provided evidence that probiotics can treat anxiety-like symptoms in a mouse model of colitis — but only if this nerve remains intact and functional[9].

The brain can also affect the microbiota and potentially contribute to the onset or exacerbation of IBS. Cryan, Bercik and others have used mouse models to show that early-life stress can change the composition of gut microbial communities, although they are yet to work out how. "Many patients who are under stress will experience diarrhoea or a change in their bowel habits as a consequence, and this could potentially affect their microbiotal composition," says Bercik.

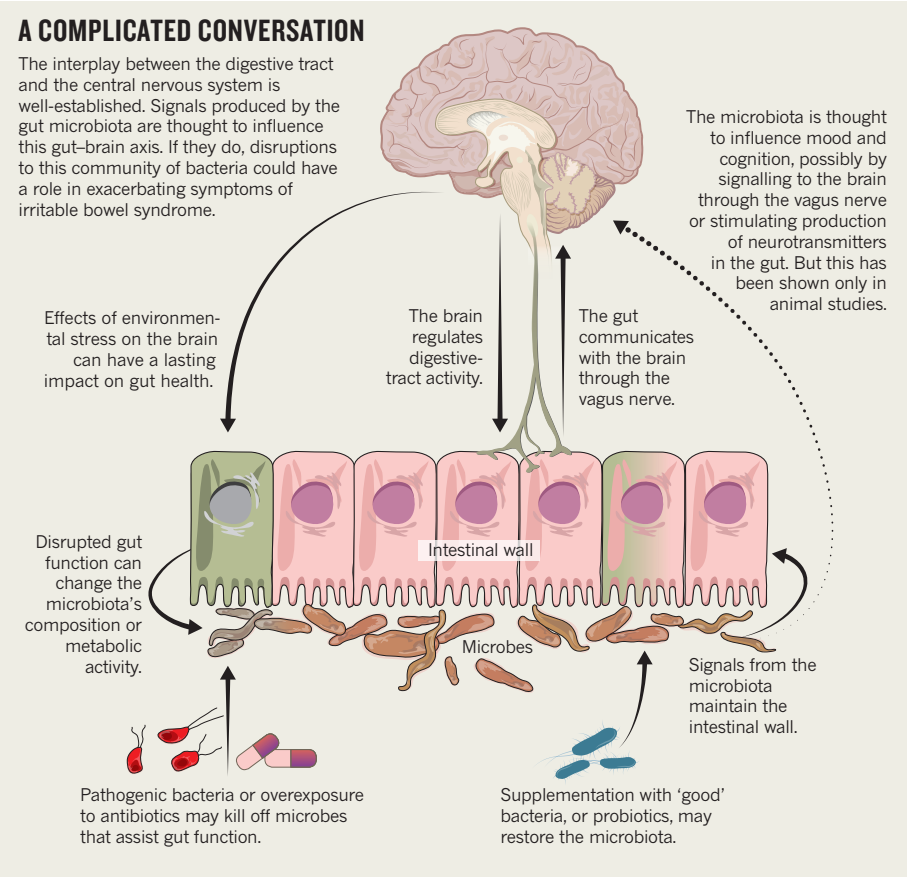
Although IBS is not recognized as an inflammatory disorder per se, there is evidence that inflammatory processes might be hyper-responsive in people with IBS, which offers another route by which the CNS might modulate the microbiome. "Depression is well known to be a pro-inflammatory state," says Eamonn Quigley, a clinical gastroenterologist at Houston Methodist Hospital in Texas. "We know the brain can influence the immune system in general, which can definitely affect the microbiome." But he stops short of labelling this as a causal mechanism by which the CNS influences IBS symptoms.

A UNIFYING THEORY

Collectively, these data suggest that IBS could arise from disturbances of one or more physiological processes, which reinforce each other to exacerbate the disease state. "None of the previous theories were able to explain IBS fully," says Bercik. "The microbiota could be one of the unifying links." The CNS response to early-life trauma could, for instance, trigger changes in the microbiota, which might then impede intestinal function and alter signalling to the brain, further amplifying the stress response. Or perhaps a gastrointestinal infection or inflammatory response to a food allergy could trigger a similar feedback cycle. "It's kind of ridiculous to pin it down to just one of these factors because they all fit into this big jigsaw puzzle," says Mayer.

This grand unifying theory of IBS is appealing, but is yet to garner strong experimental support. For one thing, links between the gut–brain–microbiome axis and IBS are built mainly on a foundation of animal studies, which may poorly reflect the physiological disruptions of patients. The germ-free rodents that have produced such striking data, for example, have an extreme phenotype that is never seen in humans. "We need to understand how this works with the milder

© 2016 Macmillan Publishers Limited. All rights reserved.



perturbations of the microbiota that we see in humans," says Simrén.

Many studies of stress and its effects on IBS have relied on models in which a rodent pup is separated from its mother for hours at a time in the first few weeks after birth. The resultant stress effects can last the animal's lifetime, with symptoms that resemble human IBS. "They're hypersensitive, they have hyperactive motility of the colon, stomach emptying slows and they're more anxious," says Mayer. But this maternal-separation model has not proved entirely persuasive. "It's fantastic that they look like they have IBS, but it's not the same thing," de Vos says. Indeed, given the apparent heterogeneity of IBS, many scientists agree that it seems far-fetched to imagine that any one animal model will ever tell the full story.

As a clinician, Mayer recognizes the seriousness of this gulf. "The animal data are almost hard to believe because they're so dramatic," he says, noting that, so far, no results have been faithfully replicated in a well-designed human study. One of the biggest problems is the paucity of human-microbiome data. Many studies of IBS have relied on faecal samples from small numbers of patients, who may have been inappropriately grouped together despite displaying different symptom patterns. "We need to have hundreds of patients in our studies to really see where there are differences between groups," says Simrén. Early studies also failed to account for gender differences or the confounding effects of dietary habits, which can make a huge difference to microbial composition.

Another important consideration is how the microbiome changes over time. An initial bacterial community is established at birth, but develops as a person matures. Given the apparent link between early-life events and subsequent development of IBS, researchers need to understand the potential effects of disruptions to the microbiota in childhood. "If we look at the microbiome of human IBS patients and notice only modest changes, that's not telling you what happened early on," says Cryan. "You could have changes from early life that potentially affected brain–gut signalling permanently."

There is also the possibility that the crucial element is how the microbiome functions, rather than what it comprises. This could help to explain why bacterial surveys of IBS have yielded such ambiguous results. "The same bacteria can exert different functions based on the environment and the food that we are supplying them with," says Bercik. The upshot is that it may be more informative to focus on the microbiome's biochemical output than on its composition.

Despite the uncertainties in the evidence, there are intriguing findings from patients that lend weight to the hypothesis that the microbiome's influence on gut–brain communication is relevant to IBS. A small study by Quigley and his colleagues[10] showed that people with IBS who have psychiatric symptoms can be distinguished from those without, purely by differences in their microbiota. Mayer's team is also continuing to pursue its brain-imaging work — preliminary results show that certain microbial profiles seem to correlate with changes in brain regions involved in pain perception and processing.

CAUSE AND EFFECT

What is still missing is solid evidence of gut microbiotal causality in IBS pathology. "We don't know yet what is the chicken and what is the egg," says Mayer. Tinkering with patient microbiomes could help scientists to address this question. Mayer and his University of California, Los Angeles, colleague Kirsten Tillisch have found that healthy women who consumed a dairy drink containing probiotic bacteria had altered activity in brain areas associated with perception of emotions and of pain and discomfort[11]. This is important evidence of microbiota-to-brain feedback, but the effect is yet to be investigated in patients.

In 2015, Bercik and his colleagues reported preliminary data that showed that giving probiotics to patients significantly alleviated their depression — and provided a measure of IBS relief. And numerous studies indicate that such treatments can lessen gastrointestinal symptoms more generally. "There's no question that probiotics have an effect," says Quigley. "What's more difficult to sort out is why they have an effect and which ones have what effect." De Vos, together with gastroenterologist Robert Brummer of Sweden's Örebro University, is exploring the causality issue from the other end, with a clinical trial to assess the effects of transplanting faecal bacteria from healthy donors into people with IBS.

Even the microbiome champions acknowledge that a better microbial mix will not be the answer for every patient. But the consensus is that a clearer picture of the 'good guys' and 'bad guys' of the gut could one day allow doctors to quickly dispatch appropriate reinforcements to the bacterial battlements, and break patients out of the self-destructive IBS cycle. ■

Michael Eisenstein *is a freelance science writer based in Philadelphia, Pennsylvania.*

1. Labus, J. *et al. Pain* **155,** 137–149 (2014).
2. O'Mahoney, S. M. *et al. Neuroscience* **277,** 885–901 (2014).
3. Rajilić–Stojanović, M. *et al. Gastroenterology* **141,** 1792–1801 (2011).
4. Jalanka-Tuovinen, J. *et al. Gut* **63,** 1737–1745 (2014).
5. Diaz Heijtz, R. *et al. Proc. Natl Acad. Sci. USA* **108,** 3047–3052 (2011).
6. Asano, Y. *et al. Am. J. Physiol. Gastrointest. Liver Physiol.* **303,** G1288–G1295 (2012).
7. Yano, J. M. *et al. Cell* **161,** 264–276 (2015).
8. Bravo, J. A. *et al. Proc. Natl Acad. Sci. USA* **108,** 16050–16055 (2011).
9. Bercik, P. *et al. Neurogastroenterol. Motil.* **23,** 1132–1139 (2011).
10. Jeffery, I. B. *et al. Gut* **61,** 997–1006 (2012).
11. Tillisch, K. *et al. Gastroenterology* **144,** 1394–1401 (2013).

© 2016 Macmillan Publishers Limited. All rights reserved.

PERSPECTIVE



An easier diagnosis

The latest iteration of the Rome criteria should improve diagnosis of irritable bowel syndrome, says **Brian E. Lacy**.

Irritable bowel syndrome (IBS) is one of the most common gastrointestinal conditions encountered by clinicians — globally, it afflicts around 11% of people of all races, sexes and socio-economic status[1]. It is associated with recurrent abdominal pain linked with defecation, a change in bowel habits (constipation, diarrhoea or both), as well as other symptoms that do not seem to relate to the gut.

Despite its prevalence, IBS remains frustratingly difficult to diagnose. The leading reason is the lack of a good biomarker. The presence of multiple, overlapping non-specific symptoms (abdominal pain, bloating, distension, diarrhoea and constipation), related to the inherent heterogeneity of the disorder, complicates matters. Then there is the lack of understanding by clinicians about the condition, and the hesitance of patients and doctors to accept a symptom-based diagnosis.

During the late 1980s, efforts to improve the diagnosis of IBS and to standardize the diagnostic process led to the development of the Rome criteria[2]. The third iteration, Rome III, was published in 2006 (ref. 3) and at the time was felt to represent the most up-to-date data on the diagnosis and treatment of IBS. Rome III defined IBS as recurrent abdominal pain or discomfort at least three days per month in the preceding three months, which is associated with at least two of the following: symptom improvement with defecation, symptom onset coupled with altered stool frequency, or symptom onset coupled with altered stool form.

During the past decade, however, it became clear that Rome III needed revision, partly because research improved our understanding of how IBS develops and of its underlying pathophysiological changes. But perhaps most importantly, because many clinicians found that the Rome III criteria were not as useful as they had hoped, and that the guidelines did not reflect real-world practice. For example, Rome III did not recommend basic laboratory tests and ignored the fact that abdominal pain often worsens, rather than improves, with defecation. The criteria were also criticized for being mainly designed for researchers and not for the busy clinician, who frequently encounters patients exhibiting multiple symptoms — both intestinal and non-intestinal. Health-care providers reported that some of the Rome III criteria were either vague or incorrect. And some researchers and clinicians argued that Rome III did not fully distinguish constipation-predominant IBS from chronic constipation, or diarrhoea-predominant IBS from other causes of diarrhoea.

The new Rome guidelines[4] aim to address these criticisms and to provide clearer guidance for health-care providers based on the latest clinical evidence. Published in May, Rome IV covers many functional bowel disorders, of which IBS is the most prevalent. It is the result of an exhaustive review of the literature and nearly three years of work by international specialists, including a team led by myself and Fermín Mearin of the Centro Médico Teknon in Barcelona, Spain, that focused exclusively on IBS.

Rome IV delivers several important changes in the diagnostic criteria for IBS. Gone is the term abdominal discomfort — the word discomfort is imprecise and is not present in every language. The criteria also require that abdominal pain is present at least one day per week on average during the preceding three months, up from just three days per month. This was based on a survey conducted on behalf of the Rome committee, which found that pain is the cardinal symptom of IBS[4]. These new criteria also recognize that IBS is often associated with irregular bowel habits — constipation, diarrhoea or a mix of both — and that common symptoms include bloating and distension. For a condition to be classified as IBS, it must be chronic, reflected by the addition that symptoms should have persisted for at least six months. The Rome criteria also now acknowledge the role of diagnostic tests to exclude other common conditions, such as coeliac disease, lactose intolerance and inflammatory bowel disease, which have similar symptoms to IBS.

By clarifying the language, updating the definition and including the option of laboratory testing, the new criteria should make it easier for doctors to diagnose IBS. The increased focus on abdominal pain reinforces what is reported by the clinicians who routinely care for these patients — that abdominal pain is the essential element of IBS. Rome IV will also help to distinguish people with IBS from those who have intermittent abdominal spasms or cramps, as well as from those who have chronic constipation or diarrhoea.

Tremendous advances have been made in our understanding of IBS over the past decade. Although it is a chronic disorder for many, and not yet curable, new therapeutic options have significantly improved the symptoms of many patients with IBS. The release of the Rome IV criteria should enhance clinicians' knowledge of this common disorder and allow them to more reliably diagnose it. In research laboratories too, a more refined definition will help scientists to identify and categorize people with IBS and its subtypes, improving our understanding of the complex physiology of the condition, and in turn, spur researchers to make even greater therapeutic discoveries. ■

THE ROME IV CRITERIA SHOULD ENHANCE CLINICIANS' KNOWLEDGE OF THIS DISORDER AND ALLOW THEM TO MORE RELIABLY DIAGNOSE IT.

Brian E. Lacy *is Chief of the Division of Gastroenterology & Hepatology at the Dartmouth-Hitchcock Medical Center, New Hampshire. He is also co-Editor-in-Chief of* The American Journal of Gastroenterology.
e-mail: brian.e.lacy@hitchcock.org

1. Lovell, R. M. & Ford, A. C. *Clin. Gastroenterol. Hepatol.* **10,** 712–721 (2012).
2. Thompson, W. G. *et al. Gastroenterol. Int.* **2,** 92–95 (1989).
3. Longstreth, G. F. *et al. Gastroenterology* **130,** 1480–1491 (2006).
4. Mearin, F. *et al. Gastroenterology* In the press (2016).

DARTMOUTH-HITCHCOCK MEDICAL CENTER

© 2016 Macmillan Publishers Limited. All rights reserved.



TETRA/GETTY

Although both considered healthy snacks, oranges are lower in FODMAPs than apples.

DIET

Food for thought

Many people with irritable bowel syndrome feel that they benefit from dietary interventions, but researchers still lack a full understanding of how food can inflame this disorder.

BY SARAH DEWEERDT

There is no doubt that people with irritable bowel syndrome (IBS) and the doctors who treat them are obsessed with diet. The past five years have seen a huge rise in awareness of the low-FODMAP diet, which omits particular carbohydrates found in wheat, onions and some fruits. Other patients are eager to go gluten free, eschewing a protein that is found in wheat and some other grains — a diet that is also popular outside the IBS community. "Patients have high expectations" for such diets, says Magnus Simrén, a gastroenterologist at the University of Gothenburg in Sweden.

The idea that food causes IBS symptoms has intuitive appeal to many. In surveys, 40–90% of people with IBS say that their symptoms have dietary triggers. "Patients have been telling us for many, many years that their symptoms are associated with eating a meal," says William Chey, a gastroenterologist at the University of Michigan in Ann Arbor. These people have historically had few options for reliable pharmaceutical treatments.

But we have been down this road before, and earlier interest in dietary interventions for IBS failed to have a lasting effect on patients' symptoms. Even now, many scientists caution that we do not know enough about the relationship between food and IBS. And although low-FODMAP and gluten-free diets are gaining traction, researchers say that a more rigorous, sceptical attitude is needed to ensure that the current interest in diet leads to lasting clinical gains.

Forty years ago, the prevailing theory on IBS was that symptoms were triggered by food intolerance or allergies. In the early 1980s, John Hunter, a gastroenterologist at Addenbrooke's Hospital in Cambridge, UK, reported success with a diet that excluded most common food allergens, including those found in milk, eggs and nuts.

This approach hit a dead end, however, when other clinics had difficulty replicating Hunter's results. Furthermore, numerous studies found that people with IBS had no higher prevalence of food allergies than the general population. By the 1990s, diet had largely been abandoned as a treatment for IBS, with the focus switching instead to pharmaceutical therapies.

But in the past decade, the pendulum has swung back, pushed both by patients' frustration with underwhelming drugs and by new research tools for investigating dietary triggers of symptoms. "We have a much better understanding now of all the different ways in which food might cause gastrointestinal problems," Chey says.

FOOD TRIGGERS

Working out which foods are implicated in causing IBS symptoms — let alone which ones trouble an individual patient — is tricky, however. "Food contains all sorts of different components," says Kevin Whelan, a dietician at King's College London. Any individual food item is chemically complex, he says — and foods are rarely eaten in isolation.

But researchers got a big break in the late 2000s. A team at Monash University in Melbourne, Australia, led by gastroenterologist Peter Gibson and dietician Sue Shepherd, conducted some of the most rigorous studies yet of the dietary triggers of IBS. Their work centred on FODMAPs (fermentable oligosaccharides, disaccharides, monosaccharides and polyols), which are carbohydrates that the body has difficulty breaking down or absorbing. They include: fructose, the form of sugar found in fruit; lactose, which is found in milk; fructans, which are polymers of fructose found in wheat, legumes and onions; and alcohol (see 'FODMAP food map').

Many FODMAPs had previously been implicated in IBS. "We've known that they cause problems," says Gibson. "But no one had ever put them together as a family that all have the same effect."

In one persuasive study[1], Gibson's team put people with IBS on a diet that drastically reduced their consumption of FODMAPs. Their symptoms improved. Then the researchers fed them fructose, fructans or a placebo. Symptoms returned in people given FODMAPs, but not in those who received the placebo. "They really have shown that FODMAPs can induce symptoms in patients with IBS," says Simrén.

Magnetic resonance imaging (MRI) studies of the digestive system have shown how FODMAPs cause problems. Poorly absorbed molecules such as fructose draw more water

© 2016 Macmillan Publishers Limited. All rights reserved.

into the small intestine, stretching it. Poorly digested FODMAPs, such as fructans, are fermented by gut bacteria, producing excess gas in the large intestine. These effects also occur in people without IBS, but the condition is also associated with visceral hypersensitivity, so even a small distension of the intestine can be painful.

In addition, in some people with IBS, food passes through the gut more quickly than normal. In these individuals, any excess water drawn into the small intestine will not have time to be reabsorbed in the large intestine, resulting in diarrhoea. "A significant proportion of people with IBS will benefit from reducing their FODMAP intake," says Whelan. He and other proponents of the low-FODMAP diet say that it can help up to 70% of people with IBS.

HIT THE FLAW

Not everyone is convinced, however. Some researchers argue that many of the studies purporting to show the benefits of a low-FODMAP diet have methodological flaws — difficulties that are commonly found in dietary studies.

The main problem is that people are aware of what they are eating, which means that unlike in studies of experimental drugs, it is difficult to blind participants to whether they are in the treatment or control group. This makes these studies more vulnerable to both placebo and nocebo effects (in a nocebo effect, the thought of eating a trigger food could bring on negative symptoms). "Expectations play a big role here," Simrén says.

To address this problem, Simrén's team tried to take expectations out of the equation. In 2015, they published a study[2] of 75 patients who were randomly assigned to follow either a low-FODMAP diet or conventional dietary advice for IBS, which focuses on avoiding large meals, fat, insoluble fibre, caffeine and gas-producing foods such as beans. Patients were told that both diets had a good chance of helping them but, crucially, the researchers never used the term 'FODMAP'. This was the largest randomized trial of the low-FODMAP diet, and the first to directly compare it with another active intervention.

After four weeks, there was no difference between the groups — about half of the people on each diet reported improvement in their symptoms. The researchers concluded that the low-FODMAP diet was as effective for IBS as the established diet. But Paul Moayyedi, a gastroenterologist at McMaster University in Hamilton, Canada, calls the results "completely negative", as the low-FODMAP diet offered no advantage over previous dietary advice.

"A low-FODMAP diet is not a diet for life."

Such disagreement about how to interpret dietary studies is common in the IBS community. "This is evidence that you actually have dietetic interventions that work," says gastroenterologist David Sanders of Sheffield Teaching Hospital in the United Kingdom. Both the low-FODMAP diet and the conventional IBS diet give a 50% chance of improvement without using drugs — and keep the patient in control. "It's wonderful, isn't it, to have that?" he asks.

But proponents and sceptics do agree that the low-FODMAP diet should be supervised by a health-care professional. Without proper advice, patients who give up high-FODMAP, but otherwise healthy foods, such as apples, artichokes and lentils, may struggle to have adequate fibre, vitamins and minerals in their diet. And some FODMAPs also encourage the growth of beneficial bacteria.

"The low-FODMAP diet removes important prebiotics from the diet and has the potential to cause really quite large perturbations in the balance of the gut microbiome," says Miranda Lomer, a dietician at King's College London who collaborates with Whelan.

For these reasons, avoiding all FODMAPs should only be a short-term undertaking. "A low-FODMAP diet is not a diet for life," says Gibson. And it should not be adopted lightly. But the proliferation of dietary information on the Internet means that, in practice, this does not always happen.

Peter Whorwell, a gastroenterologist at the University of Manchester, UK, estimates that half of his patients have put themselves on a low-FODMAP diet without taking medical advice. "That is quite worrying," he says (see page S112).

DIET DILEMMA

Low-FODMAP is not the only diet of interest to the IBS community. To try to relieve their symptoms, many people with IBS go gluten free — some studies have shown this to be effective in up to 70% of individuals with IBS.

But the role of gluten in IBS is unclear, and most 'gluten-free' diets are really wheat-free diets, so the culprit could be other compounds found in wheat. Some studies point to α-amylase trypsin inhibitors, or say that people may be benefiting from reduced consumption of insoluble fibre, which is common in wholegrain breads and other cereal products. In 1994, Whorwell found that fibre worsened IBS symptoms in more than half of patients. He suspects that fibre causes problems through similar mechanisms to FODMAPs, being fermented by gas-producing bacteria in the colon.

Or maybe it's all down to FODMAPs. Wheat contains fructans and is one of the most frequently eaten FODMAP-containing foods. In one study[3], Gibson's team cycled people with IBS through periods on low-FODMAP, low-gluten and high-gluten diets. Their symptoms diminished only during the low-FODMAP phase. "We could find no evidence of any specific response to gluten," he says.

Whatever the reason for its success, the wider popularity of the gluten-free diet, particularly in the United States and Australia, makes it clinically appealing. "It's easier to buy gluten free, it's less of a restrictive diet, and it doesn't require as much dietetic support," says Sanders. He adds that some of his patients who improve on a low-FODMAP diet eventually go gluten free instead because it's easier.

Eventually, scientists hope to identify biomarkers in the blood or the gut microbiota, or possibly both, to determine which dietary factors a particular patient is sensitive to, and what dietary changes might ease their symptoms. But such an analysis is probably years away. In the meantime, those who treat IBS tend towards pragmatism. Even Moayyedi, who is sceptical of the low-FODMAP diet, recommends that some of his patients give it a go. He says that it helps about one-quarter of the time.

Simrén says that despite little being known about the effects of gluten, he does not argue when patients say that gluten-free diets help. "I tell them, 'I don't quite know why you are better, but I'm happy that you are'," he says. "And if you feel that much better, you can continue." ■

Sarah DeWeerdt *is a freelance science writer based in Seattle, Washington.*

1. Shepherd, S. J., Parker, F. C., Muir, J. G. & Gibson, P. R. *Clin. Gastroenterol. Hepatol.* **6,** 765–771 (2008).
2. Böhn, L. *et al. Gastroenterology* **149,** 1399–1407 (2015).
3. Biesiekierski, J. R. *et al. Gastroenterology* **145,** 320–328 (2013).

FODMAP FOOD MAP

Some of these foods are prohibited on a diet that is low in the indigestible carbohydrates known as FODMAPs, but which ones? (Answer in the margin.)


Chicken


Onion


Bananas


Wholegrain bread


Cow's milk


Cabbage

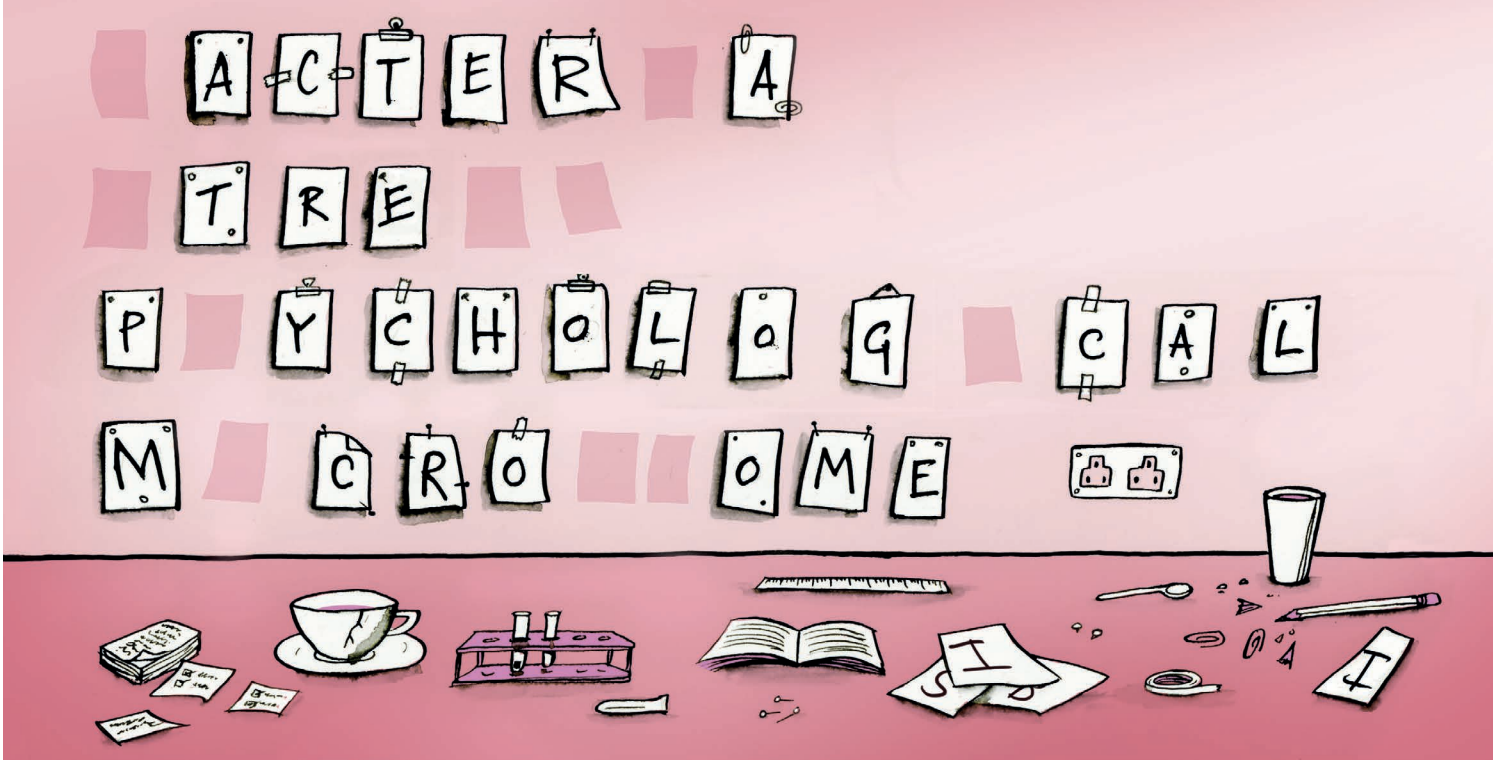

Pears


Hard cheese

ANSWER: ONION, WHOLEGRAIN BREAD, COW'S MILK, CABBAGE, PEARS.

CLOCKWISE FROM TOP LEFT: BRAND X/MARS; PHOTOALTO/MARS; COLLECTION:WHITEWISH/ISTOCK/GETTY; GETTY/MARS; OLGA POPOVA/ISTOCKPHOTO/THINKSTOCK/GETTY/MARS; COMSTOCK/MARS; GETTY/MARS; GETTY/MARS

© 2016 Macmillan Publishers Limited. All rights reserved.



DIAGNOSTICS

Filling in the missing pieces

For years, researchers have tried to find biomarkers that could aid the diagnosis of irritable bowel syndrome, and point to its underlying causes. But will one test ever be enough?

BY ALLA KATSNELSON

It was a hunch almost two decades ago that launched Mark Pimentel's interest in irritable bowel syndrome (IBS). Back then, despite the prevalence of the condition, people with IBS were often told that their symptoms were all in their heads. But Pimentel, then a gastroenterology fellow at Cedars-Sinai, a medical center in Los Angeles, California, focused elsewhere in their bodies, on the rumblings in the abdomen: bloating, distension and flatulence — in short, gas.

Pimentel wondered whether bacteria fermenting food in the gut were producing an excess of gas. In 2000, along with his mentor, Henry Lin, Pimentel reported that almost 80% of people with IBS harboured an overgrowth of bacteria in their small intestines[1]. For about half of these patients, eradicating the extra bacteria improved symptoms. Around that time, other groups were also making bacterial connections — linking food poisoning to IBS. In the same vein, when Pimentel and his colleagues infected rats with the bacterium *Campylobacter jejuni*, a common cause of travellers' diarrhoea, they found that the animals developed not only IBS-like symptoms, but also similar bacterial overgrowth[2]. That's where things began to get interesting, says Pimentel, who now directs the gastrointestinal-motility programme at Cedars-Sinai. "From that point on, we said, 'OK, something in these bugs is causing IBS.'"

The culprit, Pimentel says, is cytolethal distending toxin B (CdtB), a key part of a toxin produced by pathogens that cause gastroenteritis. Last year, Pimentel reported that in a model of IBS, rats infected with *C. jejuni* formed antibodies to CdtB, which in turn triggered an autoimmune response against cells of the gut wall[3]. Furthermore, he found that blood-plasma levels of CdtB antibodies could help to distinguish people with diarrhoea-predominant IBS (IBS-D) from those with inflammatory bowel disease or coeliac disease and from healthy controls[4]. With these results in hand, Pimentel and diagnostics company Commonwealth Laboratories of Salem, Massachusetts, developed a blood test called IBSchek. "We've never had a biomarker that says, if you're positive, you have IBS," Pimentel says. "This is that test."

People with IBS need a diagnostic test. Patients get the label only when diseases that have similar symptoms, such as Crohn's disease, coeliac disease or colorectal cancer, are excluded. The current system can leave people in diagnostic limbo — the average time to diagnosis is 6.6 years[5], during which they shuttle between doctors and undergo multiple tests, and still end up not quite being able to say that they have a disease (see page S112). Multiple attempts have been made to identify markers that could provide a positive diagnosis, but with limited success. In 2009, a team that included Anthony Lembo, a gastroenterologist at Beth Israel Deaconess Medical Center in Boston, Massachusetts, who went on to be a co-author of the 2015 CdtB-antibody clinical study, launched a 10-biomarker blood test for IBS with pharmaceutical company Prometheus Laboratories. But the test is no longer being sold. The main problem, says Lembo, was that the markers in the panel weren't related to what caused IBS.

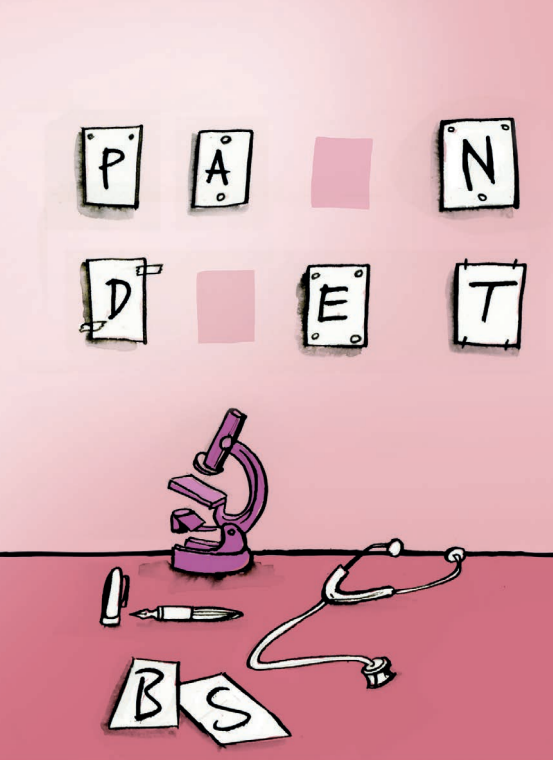
Pimentel says that his test does just that — at least for IBS-D, which about one-third of people with the condition have. But not all gastroenterologists are convinced that IBSchek is as widely applicable as its developers say, or that it represents the underlying cause of the disorder. IBS is extremely heterogeneous, and so the test requires more validation in different groups of patients, critics say. Meanwhile, some researchers are focusing on this heterogeneity to identify the underlying physiological differences that could separate IBS into distinct, diagnosable conditions.

GREAT UNCERTAINTIES

The term IBS was adopted in the 1960s, but even then clinicians struggled to define the condition (see page S102). "It had all these weird names and groupings, and there was no consensus about what these people experienced," says Lembo. Today, the condition is

SARAH J. COLEMAN

© 2016 Macmillan Publishers Limited. All rights reserved.



diagnosed on the basis of symptoms using the Rome criteria (see page S107).

The conventional diagnostic categories are functional: constipation-predominant IBS (IBS-C), IBS-D and mixed IBS. "That distinction has nested itself into clinical practice because we have drugs that work specifically for IBS-D and IBS-C," says Jan Tack, a gastroenterologist at the Univeristy of Leuven, Belgium. "But you could subdivide these people in many, many ways." For example, he says, IBS cases could be distinguished by whether they begin acutely or appear gradually, whether the predominant issue is pain or bloating, or if psychosocial symptoms such as anxiety are present. Whether and how any of these groupings align with physiological changes is not clear, but they generally warrant different treatment, says Tack.

IBS symptoms seem to be driven by a tangle of factors that include irregularities of the central nervous system, problems in gut permeability, impairment of the immune system or a microbial imbalance in the gut (see page S104). Efforts to identify biomarkers that could track IBS have been largely unsuccessful. Immune markers to detect inflammation, serotonin as a way to reveal an out-of-sync connection between the gut and the brain, and non-molecular measures such as gut hypersensitivity and brain imaging have all been explored. Even within a subgroup such as IBS-D, there seems to be a lot of variation in what might be causing the symptoms (see page S116).

"We are not at a point to tell a clinician, 'Use this test and it will solve some of your problems'."

But researchers are now going beyond IBS-C and IBS–D, albeit slowly, and identifying subtypes that have a clear relationship with the underlying biology — and hence with biomarkers, says Magnus Simrén, a gastroenterologist at the University of Gothenburg in Sweden. For instance, IBS-D can arise when bile acid, which is produced by the liver to aid digestion, fails to be absorbed in the small intestine and ends up in the colon. This bile-acid malabsorption can be identified with the SeHCAT test, which is approved in Europe, but not in the United States. The condition responds well to targeted drugs, making a precise diagnosis worthwhile. "This is the way we can go today," says Simrén, "to define subgroups of IBS patients with biomarkers."

But Pimentel contends that his IBSchek is more widely relevant. Current estimates suggest that 10–20% of IBS-D cases follow gastroenteritis, but this is based on studies that asked people to recall past incidences of stomach upset. "Two days of mild food poisoning — you may not even remember that 20 years later, now that you have IBS," Pimentel says. He suspects that 50–70% of IBS-D cases are post-infectious.

Many IBS researchers acknowledge that current estimates of post-infective IBS may be low, but they disagree that it could be as high as Pimentel's estimate. "To flip the coin completely and say that the majority must be post-infectious — we would need much better data for that," says Tack.

Pimentel's hypothesis that an antibody generated by a bacterial toxin has autoimmune properties is sound, and well-supported by animal studies, Tack says. But he is not yet convinced that this antibody causes IBS, saying that there is no independent evidence that the participants in Pimentel's 2015 study really did have post-infectious IBS. The CdtB antibody, he says, might be a kind of bystander marker: present, but not necessarily causative. People with IBS-D may have increased gut permeability, a condition that allows all kinds of macromolecules to get into the submucosal tissue more easily and trigger an antibody response — no infection required. And Simrén notes that the IBSchek antibodies were found in people who had had the condition for a long time; whether the markers can be detected when patients first present with symptoms is still to be demonstrated.

Another issue with IBSchek is that the test is negative for about half of people who do have IBS, says Lin Chang, a gastroenterologist at the University of California, Los Angeles. What's more, she says, although IBSchek was able to differentiate IBS from ulcerative colitis, it was less successful with Crohn's disease and coeliac disease. Researchers such as Chang and Tack say that wider testing is needed before large-scale commercial deployment of IBSchek. Preferably, this should involve independent researchers working in a clinical setting with large groups of new patients presenting with diarrhoea and abdominal pain, and in geographically diverse populations. "We need to see more extensive data," Tack says. "To tell a clinician now, 'Use this test and it will solve some of your problems.' I don't think we're there yet."

Pimentel dismisses this concern. The antibody study, he points out, was conducted in patients from 180 medical centres and was larger than studies of many widely used diagnostics. Moreover, he notes, there are several studies under way in different populations, and at least one, conducted by military researchers, has corroborated his findings[6].

BIG WORK

Because IBS is such a heterogeneous condition, it will inevitably take large studies with thousands of people to understand the mechanisms of the condition — and hence how to diagnose it. In June, the US National Institute of Diabetes and Digestive and Kidney Diseases will lay out such a road map in their Functional Bowel Disorders Workshop, says Chang. The pan-European network GENIEUR was launched in 2012 to identify the factors that contribute to IBS.

Beate Niesler, a molecular geneticist at the University Hospital of Heidelberg, Germany, who heads GENIEUR along with Simrén, says that the fact that each IBS drug only helps a small number of people indicates that "there are obvious sub-groupings of patients". There are several early clues that genetics and epigenetics could help to parse these groups. Associations with IBS have been found in genes related to serotonin signalling and immune, neuronal and intestinal barrier function. But most of the studies were small and the results may have been skewed by the different capabilities of each centre, Niesler says. She believes that GENIEUR, which spans 27 countries, can tease out and validate the various pathways by matching detailed data of clinical symptoms and phenotype with in-depth molecular analysis. The consortium has spent the past few years hammering out protocols, and now plans to conduct gene and RNA sequencing, analyse epigenetic markers and sequence the gut microbiome of 5,000 patients.

Almost all IBS researchers agree on the need to achieve a clearer understanding of the biological basis of this condition. This would allow patients to be given a firm diagnosis and, in turn, treatment. Pimentel is pushing towards this goal using the markers that he says reflect the underlying cause of IBS. "Now that we know that this antibody is so important," he says, the question is, "if we can get rid of it, can we cure IBS?" ■

Alla Katsnelson *is a freelance science journalist based in Northampton, Massachusetts.*

1. Pimentel, M., Chow, E. J. & Lin, H. C. *Am. J. Gastro.* **95,** 3503–3506 (2000).
2. Pimentel, M. *et al. Dig. Dis. Sci.* **53,** 982–989 (2008).
3. Pimentel, M. *et al. Dig. Dis. Sci.* **60,** 1195–1205 (2015).
4. Pimentel, M. *et al. PLoS ONE.* **13,** e0126438 (2015).
5. Drossman, D. A. *et al. J. Clin. Gastroenterol.* **43,** 541–550 (2009).
6. Pike, B. L. *et al. J. Travel Med.* **22,** 242–250 (2015).

© 2016 Macmillan Publishers Limited. All rights reserved.



GRANT WHITEHURST, UNIV. HOSP. SOUTH MANCHESTER

Q&A Peter Whorwell

A listening ear

Peter Whorwell is Director of the South Manchester Neurogastroenterology Service and a gastroenterologist at the University Hospital of South Manchester, UK. He has been researching and treating patients with irritable bowel syndrome (IBS) for more than 35 years, and has seen the depths of despair that this condition can cause.

What's it like to have IBS?

When I was a medical student in the 1970s, we were told that people with IBS were oversensitive females and that they shouldn't be taken too seriously. This erroneous perception of IBS has unfortunately persisted. But the truth is that IBS is a genuine condition that affects millions of people, men as well as women. It's actually quite common: about 15% of adults in the United Kingdom have it. Typically, patients have abdominal pain, bloating and bowel problems: some people are constipated, others have diarrhoea, some flit between the two. But they may also experience backache, fatigue or nausea, and have bladder and/or gynaecological problems. For some people, the symptoms are manageable, but for others they can be absolutely devastating.

I see the latter, patients with very severe IBS. Half of the people who come to my clinic have faecal incontinence. Some nights, they wake up in a pool of faeces. Women can have incontinence during intercourse. Sex is painful. Sometimes the bloating is so bad that they look heavily pregnant and the pain is so severe that they liken it to labour, but at least labour comes to an end. Over time, their sense of their own femininity is eroded. For some people, the symptoms are so severe and restrictive that they can't leave the house. The sick days rack up and employers are unsympathetic. Some of my patients who also have cancer tell me that having IBS is worse. People with IBS can feel totally hopeless — and still they are told that their symptoms are all in their head.

What happens when someone with these symptoms goes to see their doctor?

The first interaction is absolutely crucial. It can set a patient on a good or a bad journey through the health-care system. What I have observed in the United Kingdom and many other countries

© 2016 Macmillan Publishers Limited. All rights reserved.

is that, because general practitioners (GPs) are pushed for time, when someone with IBS describes their range of symptoms the doctor tends to focus on the one that is most troubling. If it's bladder problems, for example, they might prescribe antibiotics. If it's menstrual problems, they might send them to a gynaecologist. The result is that the IBS goes unrecognized, and treatments given to ease these other symptoms can sometimes backfire. Antibiotics, for example, can make IBS worse. Surgical interventions can aggravate the already hypersensitive gastrointestinal tract.

In the worst cases, patients end up ricocheting between their GP and various specialists. A person can end up having multiple scans, colonoscopies and gastroscopies, all of which help to exclude serious diseases such as an ulcer, Crohn's disease or cancer, but which leave the patient no closer to a definitive diagnosis or to having their symptoms eased. At this point, they are often told either that there is nothing wrong with them, or that it's probably IBS and that they need to learn to live with it. It's only after a lot of back and forth that patients eventually end up seeing someone like me, who specializes in severe cases of IBS.

Why is it so difficult to get a definitive diagnosis of IBS?

Doctors are programmed to do tests and to base their diagnoses on the results. But there is no test for IBS. It is a collection of symptoms — a syndrome, not a disease (see page S110). I don't do tests. I listen. I sit there and I let people talk. Then I piece together their symptoms and make a clinical judgement. Not everyone has the confidence to do this.

Sometimes people with IBS do receive a diagnosis from a GP or a specialist, but it's rarely unequivocal and the treatments that follow are not always successful. For example, my colleague Carol Francis, who now works at the Countess of Chester Hospital, and I showed that fibre-rich diets make only 10% of people with IBS feel better; 55% felt worse (C. Y. Francis & P. J. Whorwell *Lancet* **344,** 39–40; 1994). Yet half of all doctors still preach the myth that people with IBS should eat more fibre. Sometimes medication is prescribed. Antispasmodics can alleviate pain by relaxing the bowel muscle, and laxatives and antidiarrhoea medication can be used to control bowel symptoms. Old-fashioned tricyclic antidepressants, which also have an analgesic effect, can help, but by the time they are offered, people with IBS are so used to hearing that their illness is all in their head that often they don't want to take them.

How do you help people who have severe IBS?

I am a tertiary-care specialist. GPs and secondary-care specialists send me people who have extreme and persistent IBS. I see about 2,000 patients a year. Many of them will have been stuck in the health-care system for 5 to 10 years. These are the people who are affected so badly that they are ceasing to function socially and in the workplace; 38% of them have even contemplated suicide (V. Miller *et al. Clin. Gastroenterol. Hepatol.* **2,** 1064–1068; 2004).

By the time they find me, they are desperate for a diagnosis, which I can usually give them in our first session. I tell them I cannot cure them, but that I can help them if they help me by being engaged and involved.

I start by putting them on a very strict diet. The FODMAP diet, which excludes certain non-digestible carbohydrates, is currently very popular for IBS (see page S108). I prescribe a diet that's even more extreme and restrictive than FODMAP. Under my guidance, the patient gradually reintroduces foods to see what can be tolerated. It's an approach that helps about half of my patients.

"For some people, the symptoms are so severe and restrictive that they can't leave the house."

Next, I look at medication. I try different drugs, doses and combinations. If I suggest a tricyclic antidepressant, I explain that the low dose being offered is not for depression — most of my patients are not depressed, which is amazing when you think how awful their symptoms are — but to reduce pain. Amitriptyline, for example, helps more than half of my patients.

Can hypnotherapy help?

Yes, it can. That's why we offer this treatment after tinkering with prescriptions and diet. I have a team of specialized hypnotherapists who work with me. Last year, we reported that of 1,000 people with hard-to-treat IBS, 76% experienced a positive effect with hypnotherapy, with symptom severity reduced by half (V. Miller *et al. Aliment. Pharmacol. Ther.* **41,** 844–855; 2015). This study adds to a catalogue of more than two dozen papers that we have published showing that hypnotherapy has a positive effect for people with IBS.

Doesn't the use of hypnotherapy reinforce the idea that IBS is all in the mind?

Not at all. This isn't some sort of mind-control stage hypnotism. Gut-focused hypnotherapy, as we call it, is about inducing a state of deep relaxation and teaching people to take control of their very real, physiological symptoms. Hypnosis has been shown to reduce symptoms such as gastrointestinal sensitivity, the strength of muscular contractions in the gut and gastric-acid secretion in people with IBS. We use imagery. If someone feels bloated, we might ask them, under hypnosis, to imagine a balloon being deflated. If someone is constipated, we might ask them to visualize a river flowing slowly and then speed it up. The hypnotherapists at the centre leave patients with post-hypnotic suggestions. If the patients start to feel pain, for example, the simple act of laying their own hands on their belly should help to ease it. We're even starting to give hypnotherapy through Skype now.

Does this type of alternative therapy have any drawbacks?

As the technique becomes better known, more people are offering it as a treatment, and they sometimes do so without any qualifications or knowledge about IBS. That worries me. Some even claim to have been trained by me, but I've never met them. We're giving hypnotherapy a tremendous reputation that charlatans can live off.

What other treatments for IBS are there in the pipeline?

We're hoping to run a trial for faecal microbial transplants in the neurogastroenterology unit next year. Some people with IBS are thought to have an imbalance in their gut microbiota (see page S104). Wipe out that community of microbes and replace it with a healthy cocktail of organisms, the argument goes, and IBS may improve. In the United States, there are a few anecdotal reports of people with IBS whose symptoms have improved after faecal transplants, so we think it is worth trying. We're still very much at the formative stage and have yet to apply for ethical approval. My fear, however, is that we may be rejected because regulators regard IBS as a trivial disorder.

What's your assessment of the future of understanding and treating IBS?

I'm very gloomy about the future. IBS is the most common gastrointestinal disorder, and yet it is still heavily stigmatized. Patients feel that they get brushed off by the medical community, which considers IBS to be a mild, non-life-threatening condition. Research funding is hard to come by, and regulators are reticent to fund trials for therapies, because it is 'only IBS'. But IBS does kill people; they take their own lives. That it's only IBS is a dangerous misperception that has to change.

My hospital has dozens of respiratory physicians, cardiologists and all sorts of other specialists, but just one IBS specialist — me. I worry about what will happen when I, and others like me, retire. Two of the five hypnotherapists I work with have retired and are not being replaced. Young gastroenterologists don't want to do what I do — they want to investigate people with endoscopes. What I do mainly involves listening to people and encouraging them, by a process of trial and error, to manage their condition. I am usually able to help them — help them to leave the house, help them to return to work and help them to get their life back. I hope that this kind of approach is able to continue in the future, because it is amazing how often it works. ■

INTERVIEW BY HELEN PILCHER

This interview has been edited for length and clarity.

© 2016 Macmillan Publishers Limited. All rights reserved.



Svartediket reservoir (right) was the source of a 2004 giardiasis outbreak in Bergen, Norway.

KNUT-ARNE WENSAAS

INFECTIOUS DISEASE

Something in the water

Infections have long been thought to cause irritable bowel syndrome, but it has not been clear how. It seems that pathogens might be changing the behaviour of our gut microbiota.

BY SUJATA GUPTA

In October 2004, Guri Rørtveit, a primary-care physician in Bergen, Norway, was concerned about a noticeable rise in the number of patients complaining of persistent diarrhoea, flatulence and nausea. She began discussing the problem with colleagues over lunch breaks, but it wasn't until she got a call from the doctor across the road who had patients with the same symptoms that she really began to worry.

That doctor, Knut-Arne Wensaas, told Rørtveit that one of his patients had tested positive for the protozoan parasite *Giardia lamblia*. People become sick with giardiasis, as the resulting disease is called, when they consume contaminated food or water. A common cause of travellers' diarrhoea, *Giardia* is rare in developed countries such as Norway, where it is almost always related to travel overseas.

But Wensaas' patient had not left the country or done anything else to put herself at risk. Concerned, Wensaas and Rørtveit asked the other nine doctors in their practices to request microscopic analyses of stool samples from all of their patients with diarrhoea. Realizing that something was amiss, the lab workers who processed the requests at the Haukeland University Hospital began testing other stool samples to see if these also contained the parasite.

On 29 October, the hospital notified city officials of a giardiasis outbreak. Within days the source was traced to one of the city's main reservoirs, located some 3 kilometres from Rørtveit's office. Local residents were advised to boil their tap water, and the outbreak was finally contained.

About 1,300 people were diagnosed with giardiasis, and there were another 1,200 unconfirmed cases. In about half of these 2,500 people, the symptoms persisted for months, sometimes after repeated treatment, even though tests showed that the parasite was gone. "We knew then that there was something else going on," says Rørtveit, who is also a family medicine researcher at the University of Bergen.

The Bergen case added to a body of work showing that infections can trigger a form of irritable bowel syndrome known as post-infective (PI) IBS, which represents 6–17% of all IBS cases. There are several hypotheses about how this happens, but it seems that infections can lead to long-term changes in the gut. Understanding how IBS develops could provide clues to how some cases might be treated, or even prevented.

LINGERING ILLNESS

Even before IBS had a name, it was suspected that infection could trigger prolonged gastrointestinal problems (page S102). In his 1918 second edition of *Medical Diseases of the War*, British physician Arthur Hurst wrote: "Patients who have recovered from an acute attack of dysentery frequently remain unfit for a considerable period, which may even extend to years."

Over the next few decades, evidence slowly accrued for a link between infection and

© 2016 Macmillan Publishers Limited. All rights reserved.

STOCKTREK IMAGES/ALAMY

chronic bowel problems. Then, in 1962, Sidney Truelove and Nazir Chaudhary of the University of Oxford, UK, studied 130 patients with irritable colon syndrome, as IBS was then called. They found that one-quarter of patients had developed chronic bowel problems after a bout of gastroenteritis (which many acquired while they were serving overseas during the Second World War).

This finding was largely ignored until the 1990s, when several researchers started investigating the phenomenon in earnest. Nick Read and Kok Ann Gwee, both then at the University of Sheffield, UK, were intrigued by the observation that those who developed PI-IBS were more likely to have psychological problems than those who recovered from their gastroenteritis.

Meanwhile, Robin Spiller, a gastroenterologist at the University of Nottingham, UK, was interested in how pathogens triggered PI-IBS, and whether this was distinct from other forms of the condition. Spiller's team studied 544 people who had confirmed diarrhoeal infection and found that one-quarter said that their bowels had not returned to normal after 6 months. In patients who developed IBS, the most common pathogens were bacteria from the genus *Campylobacter* (R. C. Spiller *et al. Gut* **47,** 804–811; 2000). Furthermore, rectal biopsies revealed an increase in T lymphocytes that lasted at least three months after the infection had passed — a sign of a prolonged inflammatory response. The inflammation seemed to increase the number of serotonin-containing cells in the gastrointestinal tract, which is a potential trigger for diarrhoea.

> *"IBS was supposed to be a condition where there weren't any changes in the gut."*

When Spiller's team published this work, the prevailing wisdom was that the brain drove the bowel symptoms found in IBS. But here was evidence that, at least for PI-IBS, the problem lay in the gut itself. "IBS was supposed to be a condition where there weren't any changes in the gut — and here I was finding them," Spiller says. "It certainly went against the accepted dogma of the time."

TOXIN TROUBLE

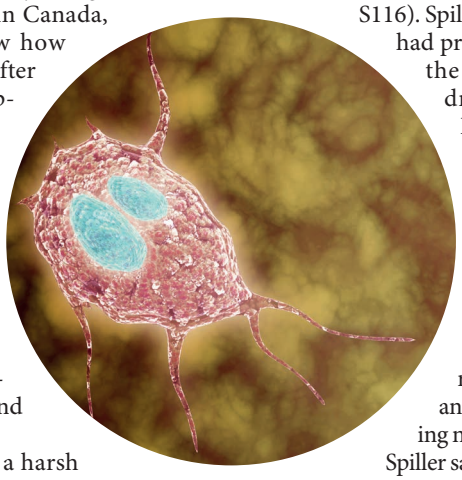
Other researchers linked PI-IBS not only with *Campylobacter* but also with *Escherichia coli*, *Salmonella* and *Shigella* bacteria, all of which produce a toxin called CdtB. Mark Pimentel, a gastroenterologist at Cedars-Sinai, a medical centre in Los Angeles, California, showed that CdtB cues the development of antibodies that in turn cause an autoimmune response. This alters gut motility and initiates the symptoms common in people with IBS. Pimentel's work suggests that in up to half of these patients, CdtB also leads to an overgrowth of bacteria in the small intestine.

But *Giardia* does not produce CdtB, making the situation in Bergen all the more mysterious. Nevertheless, the parasite somehow triggers the same symptoms as CdtB seems to, as Rørtveit and her team have shown (K.-A. Wensaas *et al. Gut* **61,** 214–219; 2011). For each confirmed giardiasis case, they identified two uninfected local people matched by age and sex but otherwise randomly selected, and found that 46% of patients still had chronic bowel problems 3 years after infection, compared with just 14% in the control group.

"It's not clear how *Giardia* might trigger IBS," Pimentel says. "It throws me off a little."

ZOMBIE BACTERIA

Andre Buret, a pathophysiologist at the University of Calgary in Canada, thinks he might know how *Giardia* triggers IBS. After hearing about what happened in Norway, he decided to see whether pathogens such as *Giardia* alter the body's microbiota. He used a technique that he and his team had developed several years earlier that allows researchers to grow a human gut microbiota outside the body and watch what happens.



***Giardia lamblia*, the parasite responsible for giardiasis.**

Gut bacteria inhabit a harsh environment that is subject to constant flushing and exposure to foreign substances. To survive, the bacteria create communities known as biofilms that are coated with protective polysaccharides. Buret wondered whether some pathogens might modify these bacterial films, inducing abnormalities that persist after the pathogen is gone. To test this idea, he put some microbial samples in a dish and infected them with *Campylobacter* and *Giardia*.

Buret's team found that *Giardia* and *Campylobacter* disintegrate the protective biofilms, so the once-beneficial bacteria can move to other parts of the gut where they become disruptive. Buret has also shown that these pathogens can alter the gene expression of freed gut bacteria, making them toxic. This interaction, he says, "transforms them into pathobionts" — the bacterial equivalent of zombies.

Like their human equivalents, zombie bacteria wreak havoc. In an unpublished study on germ-free mice, Buret found that these pathobionts target and kill the cells of the small intestine. Long after the instigating pathogen has gone, the actions of the pathobionts continue to cause gut inflammation. And with no way to get rid of the zombie menace, the inflammation can persist indefinitely.

MILITARY MIGHT

Cases of PI-IBS are particularly common in members of the armed forces who have served overseas. "Within the military, this is an underappreciated burden of disease," says Mark Riddle, a vaccine researcher at the Naval Medical Research Center in Silver Spring, Maryland.

Military service exposes people to not one but two major risk factors. First, many of those who serve overseas develop travellers' diarrhoea, which is one of the main causes of illness and lost duty days. And second, as the Sheffield researchers confirmed back in the 1990s, people are at greater risk of developing IBS if they have psychological problems — and many soldiers have post-traumatic stress disorder.

This knowledge is already affecting the hunt for treatments for PI-IBS (see page S116). Spiller and his team have had promising results with the anti-inflammatory drug mesalazine (C. Lam *et al. Efficacy Mech. Eval.* http://dx.doi.org/10.3310/eme02020; 2015). They gave the drug to 136 people with IBS. It had no effect overall, but the 13 people with PI-IBS reported less diarrhoea and cramping. The finding needs to be confirmed, Spiller says, but it supports the idea that the symptoms of IBS may arise from persistent activation of the immune system.

In 2015, Pimentel's work led to the approval by the US Food and Drug Administration of rifaximin — an antibiotic that was already used to treat travellers' diarrhoea — as a treatment for PI-IBS. Pimentel has also developed a blood test that looks for the presence of CdtB, and says that this could be an important step towards finding a cure.

Riddle has been trying to take things one step further with a preventive vaccine, and has conducted several small-scale trials of promising candidates. Volunteers receive either a placebo or a trial vaccine, and then consume a drink that contains a pathogen. He intends to test promising candidates on travellers within the next five years. Riddle also notes that those who do develop diarrhoea are immediately treated with antibiotics — and no one has gone on to develop IBS.

Despite the apparent success of antibiotic treatments, many researchers are cautious about using them to treat PI-IBS. When the microbiota have already been ravaged by infection, is it helpful to use drugs that further target the fragile intestinal ecosystem? "Knowing how precious one's microbiome is," says Spiller, "I would be very reluctant to disturb it." ■

Sujata Gupta *is a freelance science writer based in Burlington, Vermont.*

© 2016 Macmillan Publishers Limited. All rights reserved.



SARAH J. COLEMAN

DRUG DEVELOPMENT

A healthy pipeline

Drugs for irritable bowel syndrome have so far been limited, but a promising stream of options could soon enter the market.

BY BRANWEN MORGAN

After decades without any specific drugs for irritable bowel syndrome (IBS), the first few years of the millennium brought hope in the form of two approvals. But this new era turned out to be a false dawn.

In February 2000, alosetron became the first drug to be approved by the US Food and Drug Administration (FDA) for use in women who had IBS with diarrhoea (IBS-D). But within a few months there were reports of serious complications among people who were taking the drug and several deaths. Alosetron was withdrawn that November.

In 2002, the FDA approved tegaserod for IBS with constipation (IBS-C). But this drug also encountered problems. By 2007, it too had been withdrawn, this time over concerns of an increase in risk of cardiovascular problems. IBS drug development was not going well. "There was a general waning of enthusiasm on the part of pharma," says Alex Ford, a gastroenterologist at the University of Leeds, UK, "because of the bad experience with alosetron and tegaserod".

Enthusiasm was also dampened by an incomplete understanding of the multiple mechanisms that drive IBS. It is not one disease, but rather a condition comprised of a constellation of symptoms, and its diagnosis depends on the exclusion of other possible causes (see page S110). Localized gut problems include altered bowel function, bloating and abdominal pain, and for many people, anxiety and stress can cause these symptoms to wax and wane. It is difficult to stratify patients for clinical trials, to develop appropriate animal models or to identify genuine drug targets.

But hundreds of millions of people — around 11% of the global population (see 'Global variety') — have IBS, so there is a considerable need for treatments. And slowly the hurdles are falling. The FDA has approved three drugs for IBS within the past four years, and there is a bulging drug pipeline ahead with numerous compounds — some of which are already approved for other conditions — that target all manner of mechanisms from gut neurotransmission to faecal composition.

THE SEROTONIN STORY

Because people often have either IBS-D or IBS-C, the treatment mainstays have historically been antidiarrhoeal agents and laxatives, which can ease some of the functional issues, but do nothing to target the underlying mechanisms. In the 1970s, physicians began reporting that certain serotonin-targeting antidepressants were effective at slowing down bowel movement, even at low doses. And, as the evidence mounted, drug developers started to look to the serotonin system for ideas.

The gastrointestinal tract produces about 95% of the body's serotonin; this neurotransmitter is involved in gut motility, intestinal secretion and visceral sensitivity — all crucial elements of IBS symptoms. "Selective serotonin re-uptake inhibitors and tricyclic antidepressants have been effective in treating global IBS symptoms," says Baharak Moshiree, a gastroenterologist at the University of Miami in Florida.

There are seven main classes of serotonin receptors, some of which also have subtypes or variants, making for a complicated picture. The two classes most relevant to the gastrointestinal tract are 5-HT_3 and 5-HT_4. Alosetron is an antagonist of 5-HT_3 — it blocks the receptor's action and so slows the passage of stool through the gut. But in many ways it works too well. "Alosetron was a case of classical receptor pharmacology where you make a drug that binds to the one receptor you want to target — and it binds really tightly and really well," says neuroscientist Paul Bertrand at RMIT University in Melbourne, Australia. It binds so tightly that long-term use will block the receptor indefinitely. "This was obviously its downfall because it stayed there and many people became overly constipated," Bertrand says. Alosetron was made available again in 2002, but only for women with severe, uncontrollable IBS-D, and with tight restrictions — which were eased earlier this year.

Tegaserod is a 5-HT_4 receptor agonist — rather than blocking the receptor, the drug activates it and increases the signals sent to the circular muscles that wrap around the colon, speeding transit. Bertrand says that tegaserod was meant to be safer than alosetron because it is a partial agonist, so it doesn't elicit a full response from the receptor. "Its effect on the heart was unexpected," he says. "We now know it has some affinity to other serotonin receptors."

Serotonin receptors are found throughout the central and peripheral nervous system, and so there is a risk of adverse events from systemic exposure to serotonergic drugs.

© 2016 Macmillan Publishers Limited. All rights reserved.

Women seem to have more of some types of serotonin receptor than men, so men may have a different clinical response. Despite the early setbacks, serotonin remains one of the most attractive targets for IBS drug development. "The overactivity or underactivity of the enteric nervous system is really the basis of IBS as far as we understand it," explains Bertrand.

The current IBS drug pipeline includes several 5-HT_4 agonists for use in patients with IBS-C, but with one crucial difference from tegaserod. "The greater specificity of the newer drugs suggest they'll be much safer," says gastroenterologist Michael Camilleri at the Mayo Clinic in Rochester, Minnesota, and president of the American Gastroenterological Association. One such agent is prucalopride, which is highly specific for the 5-HT_4 receptor and has so far not been associated with any heart issues. It was first approved in Europe in 2009 for chronic constipation in women, and extended to use in men in 2015. It is available in Canada, but it is not yet approved in the United States. Clinical trials to extend its use to IBS are ongoing.

Alternatives to alosetron are also in the pipeline. The anti-nausea drug ramosetron, which was originally approved in East Asia in the 1990s for use in people with cancer, is also a 5-HT_3 antagonist, but it is much more potent than alosetron and so can be given at lower doses. In 2008, its use in Japan was extended to men with IBS-D, and so far there have been no reports of severe side effects. Although not currently approved for women, ramosetron has been shown to also improve their IBS symptoms (S. Fukudo *et al. Gastroenterology* **150,** 358–366; 2016).

GLOBAL VARIETY

The global prevalence of irritable bowel syndrome varies hugely, which may reflect access to health care or stigma surrounding the condition. The highest estimated rates are shown, but within countries the reported rates vary dramatically.



SOURCE: C. CANAVAN *ET AL. CLIN. EPIDEMIOL.* **6,** 71–80 (2014).

VARIED TARGETS

Fundamentally, IBS involves a problem with how the bowel functions. Because of this, many drugs aim to treat symptoms by changing the local environment in the gut — mainly by modifying fluid levels to affect the speed of faecal transit. Lubiprostone, which was approved by the FDA for IBS-C in 2008, activates chloride channels to increase fluid levels. Linaclotide, which was approved in the United States in 2012, targets the enzyme guanylate cyclase, and not only stimulates the release of ions and water, but also relieves pain. Diarrhoea is a common side effect, however. Plecanatide, which is still in phase III trials for IBS-C, activates the same receptor as linaclotide, but is a weaker agonist — this should reduce the chance of rebound diarrhoea.

Around one-quarter of IBS-D cases are caused by excess bile acid, which is produced by the liver to help with the digestion of fats. Bile-acid balance can be affected by changes in the gut microbiota and by immune dysregulation, which is why anti-inflammatory drugs and antibiotics, such as rifaximin (approved in 2015), may ease IBS symptoms. Problems also occur when too much bile acid is produced, or if too little is absorbed in the small intestine. Three bile-acid modulators are in early-stage clinical trials for IBS-D — colesevelam and colestipol, which absorb bile acid (and are already approved as cholesterol-lowering agents), and obeticholic acid, which reduces production of bile acid and is set to launch later this year to treat an autoimmune liver disease. Bile-acid levels can also be increased to help with IBS-C. Elobixibat has completed phase II trials for IBS-C in Europe and the United States; it partially inhibits absorption of bile acid, allowing more to pass into the colon.

"The greater specificity of the newer drugs suggest they'll be much safer."

Opioids have long been known to affect stool composition; constipation is a recognized side effect of morphine use. The μ-opioid receptor agonist loperamide is a well-established over-the-counter medication for general diarrhoea. It works by slowing down bowel contractions, which allows the intestines more time to absorb fluid, but it can also cause constipation. Eluxadoline, the second drug approved by the FDA for IBS-D last year, also stimulates the μ-opioid receptor, but at the same time dampens activity of the δ-opioid receptor. Pamela Hornby, lead scientist on the gastrointestinal discovery team at the pharmaceutical company Janssen near Philadelphia, Pennsylvania, which developed eluxadoline, explains that this contrasting action helps to "normalize gastrointestinal transit without causing rebound constipation". Adding, "it helps with pain, too".

BETTER TOOLS

One of the reasons that these locally acting drugs are popular with drug developers is that their effects are easier to see in animals. "We have quite a good model for motility and muscle physiology," says Jakub Fichna, a biochemist at the Medical University of Łódź, Poland. The symptoms that lack robust IBS models, however, are also those that have the greatest unmet medical need. For example, pain, which may or may not be associated with a change in bowel function, is not easy to quantify — and is even harder to target at a molecular level.

In an effort to meet this need, Fichna's group has adapted a mouse model of colon sensitivity to allow the team to replicate the sensation and pattern of visceral pain associated with IBS. Fichna suspects that multiple signals are involved in these sensations, involving systems found throughout the body. "We were working mostly in opioids and cannabinoids, but now we are also looking at nociception and are slowly entering into the serotonin area," he says. "Probably all of these are involved in visceral pain." The challenge will be to establish whether it is possible to isolate the actions of each of these systems in the gastrointestinal tract and, if they can be, whether activation of particular receptors relieves or exacerbates IBS symptoms.

IBS drug development has come a long way in the past few years. With so many treatments in the pipeline, now is an exciting time to be an IBS researcher, says Ford. For clinicians like Moshiree, it is a relief to finally have more treatment options available. "Patients usually know within a few weeks if a drug is going to work," she says, "and if they don't, you need something else to put them on." A better understanding of the multitude of mechanisms behind IBS will be pivotal, particularly in relation to how diet, anxiety and stress influence symptoms. Summing up the feeling within the community, Fichna says that pharmacology is not enough when you want to combat IBS because there are so many other factors involved. "There is no single target, no single magic pill." He relishes the challenges ahead. "This is part of the adventure." ■

Branwen Morgan *is a freelance science writer based in Sydney, Australia.*

© 2016 Macmillan Publishers Limited. All rights reserved.



SARAH J. COLEMAN

IRRITABLE BOWEL SYNDROME

4 BIG QUESTIONS

There is a long list of unknowns about the causes and symptomatology of irritable bowel syndrome (IBS). But the research challenges can be distilled into four key areas.

BY MICHELLE GRAYSON

	QUESTION	WHY IT MATTERS	WHAT WE KNOW	NEXT STEPS
1	***What are the underlying mechanisms of the syndrome?***	The variety of symptoms and of patient responses to treatment suggest that there are multiple underlying causes of IBS. Identifying specific mechanisms will help researchers to develop better treatments.	The gut can only react to insults in certain ways (for example, diarrhoea, constipation or cramps). But these symptoms could be the end point of several different diseases.	Researchers have already identified distinct causes of IBS, such as bile-acid malabsorption. Large-scale initiatives such as Europe's GENIEUR consortium are trying to tease out genetic and epigenetic contributions.
2	***What is the relationship between the disorder and mental health?***	For treatment purposes, it is important to know whether IBS starts in the gut and then causes psychological issues, or whether IBS symptoms have a neurological trigger.	Anxiety and depression sometimes accompany IBS. The serotonin system is implicated in the syndrome, and plays a part in neurological conditions. IBS is more common in service personnel who have post-traumatic stress disorder.	Better grouping of patients by phenotype will help researchers to pick out signals in large-scale trials. In the meantime, clinicians need to be careful not to assume that patients have mainly a psychological problem.
3	***How important is diet in driving IBS?***	Diet seems an obvious starting point for treating IBS. Care will be improved with a better understanding of which subset of patients could benefit from dietary interventions and why.	The fermentation of certain food components by microorganisms in the gut exacerbates symptoms. Diet can ease these, but there is little evidence of a causal link.	Larger, more rigorous trials will help to isolate the effect of food components. By working closer together, microbiologists, gastroenterologists and nutritionists could improve knowledge of the behaviour of the gut microbiota.
4	***Can the syndrome be definitively diagnosed?***	An IBS diagnosis is based on symptoms and on the elimination of other disorders. This is stressful for patients, frustrating for clinicians and inefficient for the health-care system.	Attempts to develop biomarkers for IBS have been disappointing. Breath tests and diagnostics that involve serum biomarkers are in development, but these have yet to translate to the clinic and many researchers remain sceptical.	Clearer patient phenotypes will help researchers to identify more specific biomarkers, which can also be used to stratify patients for clinical testing of targeted drugs.

Michelle Grayson *is senior supplements editor at* Nature.

© 2016 Macmillan Publishers Limited. All rights reserved.

DECADE IN REVIEW—FGIDS

‘Functional’ gastrointestinal disorders—a paradigm shift

Nicholas J. Talley

In the past decade we have witnessed an explosion in the quantity and quality of research in the functional gastrointestinal disorders. I discuss 10 top original research papers that, unless recent, have been highly cited, published in a high-impact journal and have probably shifted thinking in the field.

Talley, N. J. *Nat. Rev. Gastroenterol. Hepatol.* **11**, 649–650 (2014); published online 23 September 2014; doi:10.1038/nrgastro.2014.163

Post-dysenteric IBS was described by Chaudhary and Truelove in 1962—the initial clue that IBS might be an organic gut disease. Although important work has further characterized post-infectious IBS, Mearin and colleagues’ prospective study[1] from 2005 identified convincingly an increased incidence of dyspepsia, IBS and an overlap syndrome (dyspepsia–IBS) 6 and 12 months after a salmonella outbreak; patients with more severe gastroenteritis were at increased risk of developing one of these syndromes.[1] The overlap of functional gastrointestinal disorders (FGIDs) has remained mysterious but these key data suggest that infection might explain it. *Salmonella* spp. induces small-intestinal inflammation; a testable hypothesis is that perhaps the site of inflammation accounts for the symptom phenotype, for example proximal inflammation might be associated with dyspepsia, distal inflammation with IBS and more-extensive inflammation with the overlap syndrome.

> “One could make the case it is now time to drop the arguably pejorative term functional...”

How might inflammation induce the symptoms of FGIDs? And what about the mysterious extraintestinal symptoms these patients often report? Liebregts *et al.*[2] tested whether the release of proinflammatory cytokines in IBS is linked to symptoms and psychiatric comorbidity. They found that a number of cytokines were increased in IBS, including TNF and IL-6, and that lipopolysaccharide-induced TNF levels correlated with levels of anxiety.[2] These data further challenge the concept that IBS is a functional disease (a term that means no organic, metabolic or biochemical abnormalities), and suggest that gut-to-brain pathways might be one driver of psychological symptoms in IBS. Could psychological distress in IBS be relieved or even cured by a gut-based intervention that downregulates the cytokine response? Meanwhile, the observation that increased numbers of mast cells are present in IBS dates back to the 1990s, but it has been unclear what role mast cells have in the generation of pain and other symptoms, if any. In elegant experiments utilizing colonic biopsy samples from patients with IBS, Barbara *et al.*,[3] in 2004, observed a substantial increase in the number of degranulating mast cells releasing histamine and, strikingly, mast cells observed in close proximity to nerves correlated with abdominal pain in IBS. The role of blocking mast cells, and in particular histamine, in IBS is now an area of active interest.

Intense interest in the role of the gut microbiota in gastrointestinal and non-gastrointestinal disease continues, and the hypothesis that the colonic and small intestinal flora is altered in IBS is being actively explored. The study by Rajilic-Stojanovic *et al.*[4] is a technically excellent example of the current state of the art, demonstrating that although stool microbiota composition is highly individual, it can be used to discriminate IBS from health, with reduced numbers of Bifidobacteriaceae, Actinobacteria and Bacteroidetes but increased levels of Firmicutes in patients with IBS compared with healthy individuals. Problems remain, including limited sampling techniques, lack of an ability to culture most bacteria, and the fact that major differences have been observed in different studies. Regardless, bacterial manipulation might be one of the keys to curing FGIDs. If bacteria are important in symptom generation in IBS—for example by increased gas production that distends the lumen and induces bloating—would starving them work? In their landmark paper from 2008, Shepherd and colleagues[5] tested dietary fructose and fructan restriction then rechallenge in patients with IBS; 70% of patients receiving fructose, 77% receiving fructans and 79% receiving a mixture reported symptom breakthrough versus 14% of patients ingesting glucose. Other data now support the benefit of a low-FODMAP diet in IBS, which is being increasingly adopted although large randomized blinded trials are needed.

> “...a purely symptom-based classification of the FGIDs will be extinct in the next two decades...”

No drug has been shown to change the natural history of IBS (symptoms usually rapidly reoccur) until the past decade. If the gut microbiota is important in IBS, manipulating it should change the natural history of the condition at least temporarily. In 2011, Pimentel *et al.*[6] tested rifaximin, a minimally absorbed antibiotic, for 2 weeks in patients with diarrhoea-predominant and mixed IBS with 10 weeks follow-up. Overall, global IBS symptoms and, most notably, bloating improved although the therapeutic gain was modest (~10% of patients). The striking observation was that symptom improvement persisted during follow-up.[6] It is unknown why only a subgroup responded and whether a more targeted attack on the most relevant bacterial group would yield better results. The response is known not to be sustained long term and this factor is also unexplained; is this because of microbial or

© 2014 Macmillan Publishers Limited. All rights reserved

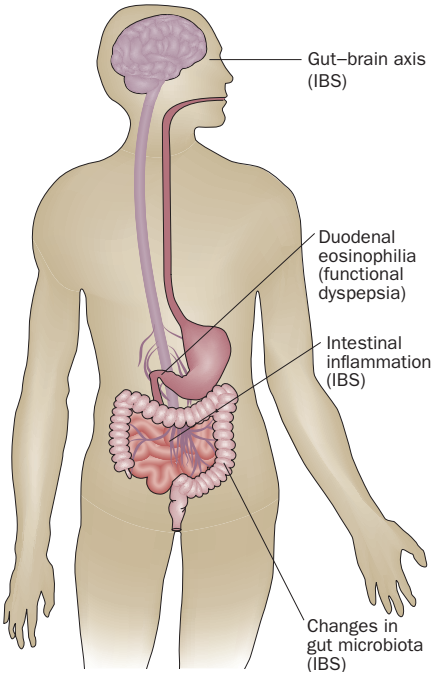


Figure 1 | Functional' gut disorders. Many patients have objective gut pathology and a gut-to-brain syndrome that explains their gastrointestinal as well as extraintestinal symptoms (e.g. anxiety, fatigue and sleep disturbance).

possibly mucosal factors? Probiotics have also shown modest symptom benefits in IBS (including a *Bifidobacteria infantis* strain) but optimal probiotic therapy remains to be defined and benefits to date are small although this area has enormous potential.[7]

At last a molecular defect in IBS has been uncovered by Coates *et al.*[8] and these results, although initially controversial, have been confirmed. Serotonin (5-HT) is released from enterochromaffin cells and mediates peristalsis and reflex responses. In this important paper, defects in serotonin signalling (including serotonin reuptake) in IBS and ulcerative colitis were observed (although numbers of enterochromaffin cells were only reduced in colitis). Gut-specific molecular alterations in serotonin might explain constipation (insufficient 5HT), diarrhoea (excess 5HT) and mixed IBS (5HT switching between excess and insufficient levels).

Finally, two key papers indicate a paradigm shift in adult functional dyspepsia. The Kalixanda study[9] was a landmark endoscopic population-based study from northern Sweden that investigated whether duodenal eosinophilia characterises 'pure' functional dyspepsia (no overlap of IBS was permitted). Patients with increased numbers of duodenal eosinophils had a nearly 12-fold increased risk of functional dyspepsia, and the pathology was linked to early satiety, a finding subsequently confirmed by multiple studies. A Belgian study[10] from 2014 not only confirmed duodenal eosinophilia in functional dyspepsia but also showed increased numbers of mast cells (arguably attributable to the majority of the study population having both dyspepsia and IBS) and more importantly impaired duodenal permeability. Increased antigen presentation (from food or bacteria) might be key in driving a subtle upper intestinal inflammatory response in predisposed patients, leading to duodenogastric reflex responses and early satiety, a condition we label as 'functional dyspepsia'.

Our concepts have changed over the past decade and the short-comings of a symptom-based only classification of FGIDs are becoming increasingly apparent (Figure 1). IBS is 'disappearing' as we identify more and more subgroups (and recognize organic diseases such as coeliac disease, microscopic colitis or bile-salt-induced diarrhoea as clinically indistinguishable from IBS). Accumulating data suggest people might have either a gut-to-brain or brain-to-gut syndrome, both with identical gut symptoms. One could make the case it is now time to drop the arguably pejorative term functional; on the basis of the latest information, many patients have gut pathology. Identifying serum and stool biomarkers to objectively diagnose IBS and functional dyspepsia now seems plausible. Prediction is dangerous but I will stick my neck out: a purely symptom-based classification of the FGIDs will be extinct in the next two decades as knowledge in this area explodes.

Faculty of Health and Medicine, University of Newcastle, HMRI Building Lot 1 Kookaburra Circuit, New Lambton Heights, NSW 2035, Australia.
nicholas.talley@newcastle.edu.au

Competing interests
The author declares no competing interests.

1. Mearin, F. *et al.* Dyspepsia and irritable bowel syndrome after a Salmonella gastroenteritis outbreak: one-year follow-up cohort study. *Gastroenterology* **129**, 98–104 (2005).
2. Liebregts, T. *et al.* Immune activation in patients with irritable bowel syndrome. *Gastroenterology* **132**, 913–920 (2007).
3. Barbara, G. *et al.* Activated mast cells in proximity to colonic nerves correlate with abdominal pain in irritable bowel syndrome. *Gastroenterology* **126**, 693–702 (2004).
4. Rajilic-Stojanovic, M. *et al.* Global and deep molecular analysis of microbiota signatures in fecal samples from patients with irritable bowel syndrome. *Gastroenterology* **141**, 1792–1801 (2011).
5. Shepherd, S. J., Parker, F. C., Muir, J. G. & Gibson, P. R. Dietary triggers of abdominal symptoms in patients with irritable bowel syndrome: randomized placebo-controlled evidence. *Clin. Gastroenterol. Hepatol.* **6**, 765–771 (2008).
6. Pimentel, M. *et al.* Rifaximin therapy for patients with irritable bowel syndrome without constipation. *N. Engl. J. Med*. **364**, 22–32 (2011).
7. O'Mahony, L. *et al. Lactobacillus* and *bifidobacterium* in irritable bowel syndrome: symptom responses and relationship to cytokine profiles. *Gastroenterology* **128**, 541–551 (2005).
8. Coates, M. D. *et al.* Molecular defects in mucosal serotonin content and decreased serotonin reuptake transporter in ulcerative colitis and irritable bowel syndrome. *Gastroenterology* **126**, 1657–1664 (2004).
9. Talley, N. J. *et al.* Non-ulcer dyspepsia and duodenal eosinophilia: an adult endoscopic population-based case-control study. *Clin. Gastroenterol. Hepatol.* **5**, 1175–1183 (2007).
10. Vanheel, H. *et al.* Impaired duodenal mucosal integrity and low-grade inflammation in functional dyspepsia. *Gut* **63**, 262–271 (2014).

© 2014 Macmillan Publishers Limited. All rights reserved

PRIMEVIEW

IRRITABLE BOWEL SYNDROME

nature REVIEWS DISEASE PRIMERS

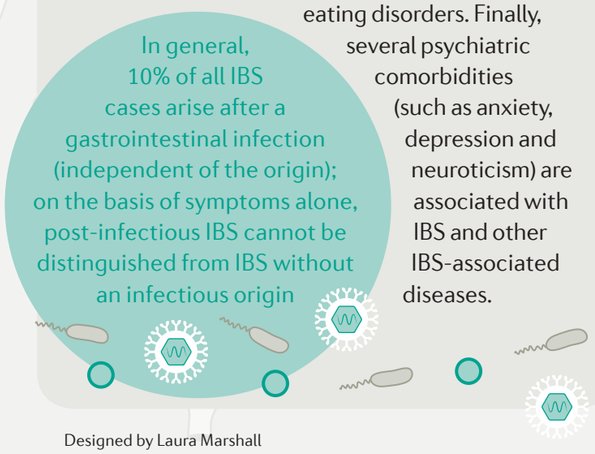
For the Primer, visit doi:10.1038/nrdp.2016.14

Irritable bowel syndrome (IBS) is a functional gastrointestinal disorder and is not associated with structural or biochemical abnormalities in the gut. Gastrointestinal symptoms include abdominal pain or discomfort, stool irregularities and bloating. In addition, IBS is often associated with somatic comorbidities (for example, pain syndromes, overactive bladder and migraine), psychiatric conditions (including depression and anxiety) and visceral sensitivity.

EPIDEMIOLOGY

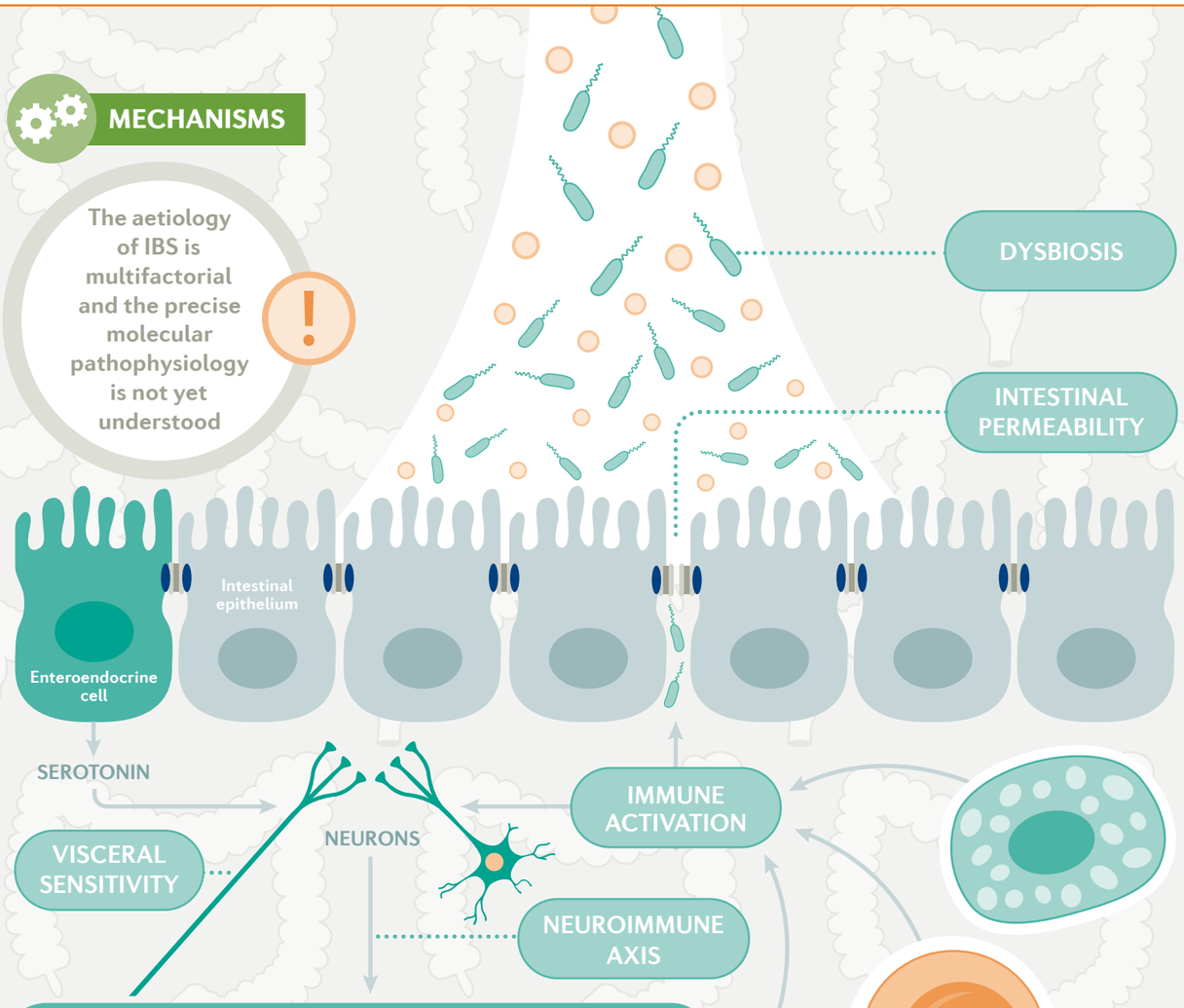
IBS has a global estimated prevalence of ~11%, with rates of 5–10% in most European countries, the United States and China. Approximately 2% of all patients show spontaneous disease remission every year. Women are more prone to developing IBS than men. IBS is strongly associated with other conditions and >20% of patients with IBS have other functional gastrointestinal disorders, including gastroesophageal reflux disease, diarrhoea, incontinence and pelvic floor dyssynergia. In addition, IBS is linked to functional, non-gastrointestinal syndromes such as pain syndromes or disorders associated with dysfunction of the autonomic nervous system, including overactive bladder, chronic fatigue syndrome and eating disorders. Finally, several psychiatric comorbidities (such as anxiety, depression and neuroticism) are associated with IBS and other IBS-associated diseases.

In general, 10% of all IBS cases arise after a gastrointestinal infection (independent of the origin); on the basis of symptoms alone, post-infectious IBS cannot be distinguished from IBS without an infectious origin



MECHANISMS

The aetiology of IBS is multifactorial and the precise molecular pathophysiology is not yet understood



The brain–gut interaction is bidirectional. Although structural and functional brain abnormalities have been described in IBS, whether they are a cause or a consequence of the gastrointestinal symptoms is unknown.

QUALITY OF LIFE

Both the gastrointestinal symptoms and the psychiatric comorbidities impair quality of life in patients with IBS, which is even lower than in those with gastroesophageal reflux disease, diabetes mellitus or severe chronic kidney disease.

OUTLOOK

Given that IBS is a heterogeneous condition, many drugs have failed in randomized controlled trials. Adequate patient stratification might reveal that some subgroups of patients benefit from available drugs, making the evaluation of new drugs more efficient.

DIAGNOSIS

IBS can be debilitating in some patients, whereas others might have only mild or moderate symptoms. IBS is diagnosed using diagnostic criteria that take a combination of symptoms into account, including abdominal pain or discomfort associated with chronic or recurring abnormal bowel habits. Other diseases (including functional gastrointestinal diseases, such as dyspepsia and gastroesophageal reflux disease) that cause typical IBS symptoms should be excluded. IBS can be associated with diarrhoea, constipation or both, and patients often switch between these different subtypes over time.

MANAGEMENT

Only half of patients with IBS seek medical treatment and pain is an important driver. A stepped-care approach to treatment is recommended, which includes psychotherapy. Medical treatment is aimed at improving bowel habits and/or visceral pain and consists of antispasmodic drugs, low-dose antidepressants, gut motility accelerants and antidiarrhoeals — depending on the symptoms.

Dietary advice plays an important part in IBS treatment because most patients have some form of food intolerance (that is, a physiological response to food allergens that is not associated with an immune response). Diets containing low levels of fermentable oligosaccharides, disaccharides, monosaccharides and polyols (low FODMAP diets) might improve symptoms, as can other dietary interventions.



Designed by Laura Marshall

Article number: 16015; doi:10.1038/nrdp.2016.15; published online 24 March 2016

© 2016 Macmillan Publishers Limited. All rights reserved.

REVIEW

Open

Intestinal Microbiota And Diet in IBS: Causes, Consequences, or Epiphenomena?

Mirjana Rajilić-Stojanović, PhD[1], Daisy M. Jonkers, PhD[2], Anne Salonen, PhD[3], Kurt Hanevik, MD, PhD[4], Jeroen Raes, PhD[5], Jonna Jalanka, PhD[6], Willem M. de Vos, PhD[3,6,7], Chaysavanh Manichanh, PhD[8], Natasa Golic, PhD[9], Paul Enck, PhD[10], Elena Philippou, PhD[11], Fuad A. Iraqi, PhD[12], Gerard Clarke, PhD[13], Robin C. Spiller, MD, PhD[14] and John Penders, PhD[15]

Irritable bowel syndrome (IBS) is a heterogeneous functional disorder with a multifactorial etiology that involves the interplay of both host and environmental factors. Among environmental factors relevant for IBS etiology, the diet stands out given that the majority of IBS patients report their symptoms to be triggered by meals or specific foods. The diet provides substrates for microbial fermentation, and, as the composition of the intestinal microbiota is disturbed in IBS patients, the link between diet, microbiota composition, and microbial fermentation products might have an essential role in IBS etiology. In this review, we summarize current evidence regarding the impact of diet and the intestinal microbiota on IBS symptoms, as well as the reported interactions between diet and the microbiota composition. On the basis of the existing data, we suggest pathways (mechanisms) by which diet components, via the microbial fermentation, could trigger IBS symptoms. Finally, this review provides recommendations for future studies that would enable elucidation of the role of diet and microbiota and how these factors may be (inter)related in the pathophysiology of IBS.

Am J Gastroenterol 2015; 110:278–287; doi:10.1038/ajg.2014.427; published online 27 January 2015

INTRODUCTION

With a prevalence of approximately 10–15%, irritable bowel syndrome (IBS) is one of the most common gastrointestinal (GI) disorders in the industrialized world (1). The high prevalence together with the reduced quality of life and associated co-morbidities in patients suffering from IBS put a significant negative burden on both patients and society (2). IBS is a heterogeneous functional disorder and commonly subtyped according to the prevailing bowel habit into IBS with constipation (IBS-C), IBS with diarrhea (IBS-D), mixed IBS with both constipation and diarrhea and unsubtyped IBS with neither constipation nor diarrhea using the Rome III criteria (3).

Although the etiology of IBS is incompletely understood, it is generally regarded as a multifactorial disorder involving the interplay of both host and environmental factors, including diet. The host factors include central factors, such as aberrant stress responses, psychiatric co-morbidity, and cognitive dysfunctions, whereas intestinal functions are also involved, e.g., dysmotility, visceral hypersensitivity, low-grade immune activation, altered barrier function, and the intestinal microbiota composition (4).

During recent years, perturbations in the intestinal microbiota are increasingly being linked to the pathophysiology of IBS. The increased risk of new onset IBS after an episode of gastroenteritis (5) and the association with prior antibiotic use (6) support the importance of the intestinal microbiota in IBS. Numerous studies have demonstrated altered microbial profiles in (specific subgroups of) IBS patients compared with healthy individuals (7–11). The intestinal microbiota strongly interacts with exogenous factors, in particular diet, which may also directly or indirectly provoke IBS symptoms.

In this review, we describe the current evidence regarding the impact of diet and the intestinal microbiota and their (inter)relation on the pathophysiology of IBS.

[1]Department of Biochemical Engineering and Biotechnology, Faculty of Technology and Metallurgy, University of Belgrade, Belgrade, Serbia; [2]Division Gastroenterology-Hepatology, School for Nutrition, Toxicology and Metabolism (NUTRIM), Maastricht University Medical Center+, Maastricht, The Netherlands; [3]Immunobiology Research Programme, Department of Bacteriology and Immunology, University of Helsinki, Helsinki, Finland; [4]Department of Clinical Science, University of Bergen, Bergen, Norway; [5]Department Microbiology and Immunology, KU Leuven, Leuven, Belgium; [6]Department of Veterinary Biosciences, Microbiology, University of Helsinki, Helsinki, Finland; [7]Laboratory of Microbiology, Wageningen University, Wageningen, The Netherlands; [8]Digestive System Research Unit, University Hospital Vall d'Hebron, Ciberehd, Barcelona, Spain; [9]Laboratory for Molecular Microbiology, Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Belgrade, Serbia; [10]Department of Psychosomatic Medicine and Psychotherapy, University Hospital Tuebingen, Tübingen, Germany; [11]Department of Life and Health Sciences, School of Sciences and Engineering, University of Nicosia, Nicosia, Cyprus; [12]Department of Clinical Microbiology and Immunology, Sackler Faculty of Medicine, Tel-Aviv University, Tel-Aviv, Israel; [13]Department of Psychiatry and Alimentary Pharmabiotic Centre, University College Cork, Cork, Ireland; [14]Nottingham Digestive Diseases Biomedical Research Unit, University of Nottingham, Queens Medical Centre, Nottingham, UK; [15]Department of Medical Microbiology, School for Nutrition, Toxicology and Metabolism (NUTRIM), Maastricht University Medical Center+, Maastricht, The Netherlands. **Correspondence:** John Penders, PhD, Department of Medical Microbiology, School for Nutrition, Toxicology and Metabolism (NUTRIM), Maastricht University Medical Center+, PO Box 5800, Maastricht 6202, AZ, The Netherlands. E-mail: j.penders@maastrichtuniversity.nl
Received 23 September 2014; accepted 9 December 2014

THE IMPACT OF DIET ON THE INTESTINAL MICROBIOTA

The fetal intestine is thought to be sterile, although some recent studies indicate that colonization may already start before delivery, by bacterial transmission through the placental barrier (12,13). The colonization during and upon delivery involves a succession of microbial populations, a seemingly chaotic process strongly influenced by birth mode and infant feeding. Vaginally born infants are initially colonized by maternal fecal and vaginal microbes, such as *Bifidobacterium*, *Bacteroides*, *Prevotella,* and *Lactobacillus* spp., whereas infants born by cesarean section are inoculated by typical skin bacteria and bacteria from the hospital's environment (14,15).

The strong impact of diet on the indigenous microbial communities can already be observed even before weaning. Compared with formula-fed infants, the microbiota of breastfed infants is less diverse and more dominated by *Bifidobacterium* spp. (16–18). This is generally attributed to unique bioactive compounds such as human milk oligosaccharides that serve as metabolic substrates for a limited number of microbes (19), although breast milk itself also harbors a transferrable microbiota (20). A rapid diversification of the infants' intestinal microbiota and shift toward an adult-like microbiota occurs with the introduction of solid foods (21,22); yet, the maturation of the microbiota appears to continue in parallel to physiological development of, for example, the gastrointestinal tract and the central nervous system, throughout childhood and adolescence (23,24).

Once established, the intestinal microbiota appears to be relatively stable within individuals over time (25), whereas high variability between individuals is observed (26–29).

Pronounced differences in the composition and functional capacity of the human microbiota between geographically distant populations point toward an important role of the diet in these inter-individual variations (26,30–32). The microbiota in non-Western populations, consuming a diet high in plant-derived carbohydrates, has consistently been shown to be more diverse and enriched in *Prevotella* spp. at the expense of *Bacteroides* spp. as compared with Western populations consuming a diet high in animal protein, sugar, starch, and fat (reviewed in (33)). It remains to be determined to what extent the microbiota composition is shaped by the diet, as the impact of genetics, ethnic, and cultural factors, such as hygiene and living conditions, cannot be ruled out in such studies comparing distinct populations.

Nevertheless, habitual long-term dietary patterns have also been directly linked to intestinal microbial enterotypes. Especially, protein and animal fat intake has been associated with the *Bacteroides* enterotype, whereas a high carbohydrate intake was associated with the *Prevotella* enterotype (34). The effect of short-term dietary interventions on the microbiota composition appears to be only modest (34,35), unless the intervention comprises an extreme switch in diet (36–38). David *et al.* (37) demonstrated that even within days after the transition to a diet entirely composed of animal or plant products, changes in microbial structure, metabolic activity, and gene expression can be observed. In subjects receiving the animal-based high fat and protein diet, the abundance of bile-tolerant bacteria (*Alistipes, Bilophila* and *Bacteroides*) together with branched short-chain fatty acids indicative of amino-acid fermentation increased, whereas levels of bacteria metabolizing dietary plant polysaccharides (*Eubacterium rectale, Roseburia, Ruminococcus bromii*) decreased. In agreement, switching to a diet high in resistant starch has been shown to increase *Ruminococcus bromii* and *Eubacterium rectale* related species (39), which are known for their saccharolytic properties. Altogether, these studies indicate that the intestinal microbial community structure as well as its function and metabolic output are influenced by our diet, especially in the case of clearly distinct dietary patterns.

INFLUENCE OF DIET ON IBS

Surveys on perceived food intolerance show that 64–89% of IBS patients report their symptoms to be triggered by meals or specific foods (40–42) The majority indicate that they limit or exclude certain foods without professional counseling (40,42), thereby increasing the risk of inadequate dietary intake. Monsbakken *et al.* (40) found indications for inadequate dietary intake in 12% of IBS patients studied, whereas others found adequate or even increased intake of some nutrients (43–45). Overall, data on habitual dietary intake are limited and may depend on the study population included. Meal-related symptom aggravation is further supported by studies showing a postprandial worsening of pain using symptom diaries for 6 weeks (46) or rectal barostat (47,48).

Foods often reported to provoke symptoms include wheat/grains, vegetables, milk products, fatty foods, spicy foods, coffee, and alcohol (40–42,49) and are especially reported to be associated with abdominal pain and gas problems (41,42). Overall, the percentage of responders on exclusion diets varies from 15 to 71% (50), but it has to be noted that most studies suffer from major methodological limitations and only a minority of subjects was found to react positively to double-blind food challenges (51–53).

Although placebo and possibly nocebo effects have to be considered in IBS, possible physiological mechanisms of (perceived) food intolerance include an exaggerated sensory and motor response and/or incomplete absorption, which may lead to symptoms in a susceptible host. High fat intake, for example, is associated with an exaggerated colonic motor response to eating and increased visceral sensitivity in IBS patients, whereas duodenal lipid infusions delay small intestinal transit. Randomized controlled trials adjusting fat intake are, however, limited (54). Poorly absorbed carbohydrates, like lactose, fructose, and galacto- or fructo-oligosaccharides (fructans), can result in luminal distension by osmotic effects and increased gas production due to microbial fermentation. Although this is a normal physiological phenomenon, it can result in symptoms in subjects with an altered microbiota, increased visceral sensitivity, and/or abnormal gas handling. A recent study by Yang *et al.* (55) found increased mucosal mast cell numbers, serum tumor necrosis factor-α, rectal sensitivity, and anxiety in lactose intolerant IBS-D patients, supporting that neuro-immune modulation of visceral function is a potential underlying mechanism. Perceived intolerance to lactose, fructose, and fructans is frequently

© 2015 by the American College of Gastroenterology

reported and uncontrolled studies point to symptom improvement after removal of milk containing products or wheat from the diet (42). However, these studies should be interpreted with caution because of a large placebo and nocebo response in IBS. The overall evidence for an increased incidence of lactose intolerance in IBS patients was found to be weak and to be moderate—weak for the benefit of low lactose intake in these subjects (56).

Evolving from previous studies on lactose and fructose intolerance, a more generalized approach of intolerance to poorly absorbed and rapidly fermented carbohydrates has been introduced: FODMAPS, including fermentable oligo-, di- and monosaccharides and polyols. Both retrospective and prospective open studies have shown reduced symptoms after introduction of a low-FODMAP diet in IBS patients with suspected or proven fructose or lactose malabsorption (57,58). These findings are confirmed by randomized controlled trials, showing a significant improvement of overall symptoms, abdominal pain and bloating, in patients with bloating and/or diarrhea (59) and in small (n = 15–30) unselected IBS patient groups (60,61). One study showed recurrence of symptoms on re-challenge with fructose and/or fructans (62). Involvement of gas production is supported by findings that IBS patients produced more hydrogen relative to healthy controls and lacked an increased methane production on a high vs. low-FODMAP diet (60). However, it should be noted that the long-term benefit of a low-FODMAP approach in large (unselected) patient groups is unclear, especially considering its invasiveness, which requires strict guidance by a dietician, and the potential risks of, for example, reduced fiber intake and changes in microbiota composition and activity (59,63). Furthermore, randomized controlled trials are needed to prove whether it is superior to other dietary interventions (e.g., lactose or fructose reduction) or general dietary advice (e.g., the NICE guidelines). Finally, the FODMAP levels vary between studies, countries, and products, which should be taken into account when interpreting study results and considering low-FODMAP diets.

Foods can also evoke symptom onset by immune activation or altered neuro-endocrine responses (42). Food allergy or intolerance associated with IgE-mediated immune responses is uncommon and evidence for IgG/IgG4-mediated hypersensitivity is inconclusive (42,64). However, in subsets of IBS patients, increased numbers of T-lymphocytes, mast cells, eosinophils, and/or enteroendocrine cells have been found (65–67), but their exact role in food intolerance and symptom generation is unclear.

Interest in the role of gluten intolerance (and the possible benefit of gluten-free diets) is increasing. A systematic meta-analysis found a pooled prevalence for celiac disease up to 4% among IBS patients (68). The prevalence of non-celiac gluten sensitivity among unselected IBS patients is unknown and complicated by clear diagnostic criteria and overlap in symptoms. Although symptom reduction after a gluten-free diet (69) and worsening of symptoms after a gluten challenge (70) have been reported, the recent study Biesiekierski *et al.* (71) found no symptom induction by giving pure gluten to IBS patients following a gluten- and FODMAP-free diet. These findings suggest that not gluten but fructans and/or other components might contribute to symptoms in perceived "wheat" intolerance. Furthermore, Carroccio *et al.* (72) showed that the majority of IBS patients diagnosed with wheat sensitivity had multiple food sensitivities.

Dietary supplementation studies focus on probiotics, prebiotics, synbiotics, and fiber intake. A recent meta-analysis on the use of probiotics shows significant effects on subjective global symptom improvement, but study heterogeneity is statistically significant and no conclusions can be drawn on which individual species or strains are most beneficial (73). Studies on prebiotics and synbiotics are limited and insufficient to draw conclusions. Evidence for the benefit of dietary fiber was found for soluble fiber intake only (74), but it should be noted that overall study quality was moderate and more studies are needed on its effects in IBS subgroups.

Although several food components are associated with symptom generation, detailed analyses on the effect of food intake on symptom generation and underlying mechanisms are limited. In addition, many studies suffer from methodological limitations, in part inherent to the complexity of food research.

MICROBIOTA ALTERATIONS IN IBS

Following the first comprehensive analysis reporting a distinctive intestinal microbiota composition in IBS patients (75), several studies have identified differences between the microbiota of IBS patients, or subgroups thereof, and healthy controls (reviewed in (7)). Recently, clinical guidance regarding the modulation of intestinal microbiota in IBS was provided by the Rome Team Working Group (76), which concluded that there is good evidence supporting the concept that the intestinal microbiota is perturbed in patients with IBS. However, despite a growing consensus regarding an association between the intestinal microbiota and IBS, results of current studies lack general consensus and a specific microbial signature in IBS remains elusive (**Figure 1**). Lack of detailed phenotypic characterization of patients, small sample sizes, and the cross-sectional study designs (providing only a single snapshot of the microbiota composition) in the majority of studies may all contribute to the fact that the markers of the IBS microbiota reported in several studies (marked in bold on **Figure 1**) are not reproducibly detected in all cohorts. Moreover, exogenous factors, including diet, are most often not taken into account, despite the demonstrated effect of diet on the microbiota and the potential alterations in diet associated with IBS. Finally, it is noteworthy that Jeffrey and colleagues identified distinct subsets of IBS patients, not corresponding to the traditional IBS subtypes, with an altered or normal-like microbiota composition (8), indicating that the disturbed microbiota might be relevant for the pathology of part of the IBS patients. Moreover, those with a normal-like microbiota had more adverse psychological factors, suggesting that central factors may predominate over microbiological factors in some but not all IBS patients.

Several recent comprehensive studies of the microbiota in IBS have reported an increase in the relative abundance of *Firmicutes*, mainly *Clostridium* cluster XIVa and *Ruminococcaceae*, together with a reduction in the relative abundance of *Bacteroidetes* (7–9,77). Bifidobacteria have been shown to be depleted in both

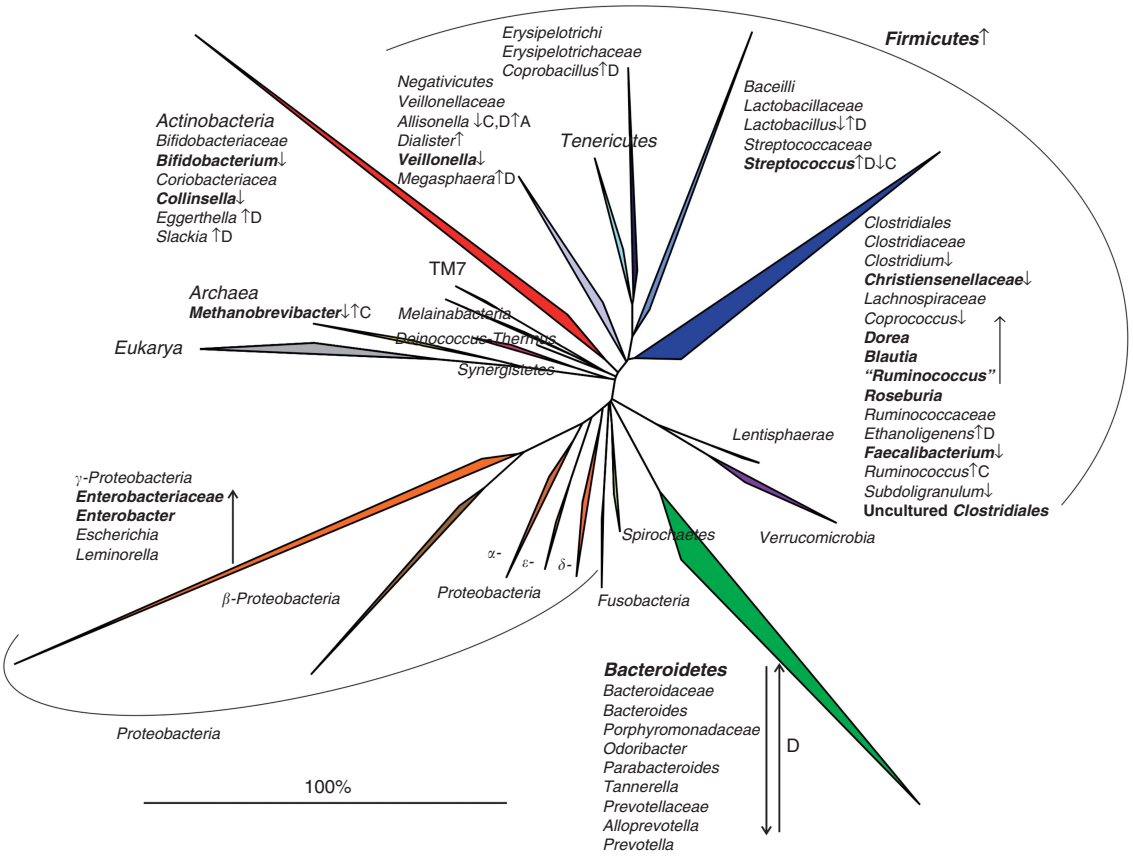


Figure 1. Phylogenetic tree of the human intestinal microbiota with indicated microbial groups that are significantly altered in irritable bowel syndrome (IBS) patients relative to controls. The increase and decrease in microbial groups in IBS patients are marked by arrows facing up and down, respectively, and in bold if reproduced by at least two independent studies. If microbial groups are altered only in diarrhea or constipation predominant IBS patients, letter D or C in superscript follows the arrows, respectively. The figure is generated on the basis of the data published in (8–11,75,77–79,82,85,98,116).

fecal and mucosal samples of IBS patients (9,78–81), which adds to the trend toward a modest beneficial effect of *Bifidobacterium* supplementation, in terms of the improvement of global IBS symptoms and pain scores as reported in a meta-analysis of probiotic trials.(73) A lower diversity and a higher instability of the (*Firmicutes* fraction or function of the) microbiota in IBS patients relative to controls have been reported (82–85), although these findings need confirmation in studies with sufficient depth of the microbiota analysis.

Most if not all studies that compare microbial profiles in IBS patients and healthy controls are associative and cannot distinguish cause from consequence. However, evidence of a potential causal role for the GI microbiota is supported by a fecal transplantation experiment in which visceral hypersensitivity, presumed to underpin abdominal pain in a subgroup of IBS patients, could be transferred via the microbiota of IBS patients to previously germ-free rats (86). Furthermore, germ-free mice colonized with the microbiota from IBS-D patients have been shown to exhibit a faster GI transit and impaired intestinal permeability as compared with mice gavaged with the microbiota from healthy humans (87). The mechanisms through which the microbiota exerts these effects are not fully understood, but the fact that supernatants from colon biopsies of IBS patients, with increased levels of histamine and proteases, excite human submucosal neurons could indicate a brain–gut axis connection (88). Furthermore, an increase in fecal serine proteases has been linked with symptom development in IBS patients (89,90). Although the origin of both histamine and proteases could be both human and microbial, it is noteworthy that some of the *Firmicutes* bacteria that were found to be increased in abundance in IBS patients (8,9) are known to secrete large amounts of extracellular proteases (91). Alternatively, the predominantly endogenous fecal proteases might not be degraded by microbes because of the accelerated transit (as seen in IBS-D patients) or because of the disturbed metabolic activity of the microbiota (92). Serine protease inhibitors, the antagonists of serine proteases, are produced by many bacteria, including bifidobacteria (93), which may contribute to the possible benefit of bifidobacterial supplementation.

Dysregulated intestinal immune function, chronic low-grade mucosal inflammation, and increased mucosal permeability and barrier dysfunction have all been suggested as putative pathogenic mechanisms in IBS, in which the intestinal microbiota might have a role (reviewed in (76,94)). Moreover, the bidirectional interactions between the intestines and the central nervous system, which have an important role in the pathogenesis of IBS,

© 2015 by the American College of Gastroenterology

have suggested to be modulated by the microbiota (reviewed in (95)). Yet, there is a lack of data on the exact mechanisms through which the host-microbiota interactions underlie pathophysiology and generate symptoms. Identifying these mechanisms is further complicated by the fact that the majority of GI microbes remain uncharacterized (96). Indeed, most putative microbial markers of IBS are among uncultured bacteria. For example, uncultured bacteria related to *Ruminococcus torques* are significantly enriched in IBS patients in several cohorts (9,11,75), and their abundance positively correlates with bowel symptoms (9,10,97). It is intriguing that this bacterial group has been shown to be suppressed following a multispecies probiotic treatment that alleviated IBS symptoms (98). Similarly, uncultured bacteria within the *Clostridiales* order are reproducibly detected in significantly depleted abundance in IBS (9,10) but also in ulcerative colitis (99). Focus on the function of these uncultured bacteria should be of major interest for future studies.

MICROBIOTA–DIET INTERACTIONS IN IBS

Intestinal microbes have an important role in the digestion of dietary components, resulting in metabolites that may directly or indirectly contribute to IBS symptoms (**Figure 2**).

Colonic fermentation of carbohydrates mainly results in short-chain fatty acids (SCFAs, including acetate, propionate, and butyrate) and variable amounts of hydrogen and carbon dioxide. Several low-abundant microbial groups can subsequently dispose the produced hydrogen into acetate, hydrogen sulfide, and methane. The fermentation of protein residues, primarily in the distal colon, produces a variety of metabolites, including ammonia, organic acids, heterocyclic amides, and phenolic and indolic compounds, that are toxic and detrimental to gut health (100). Moreover, fat and digested proteins increase the excretion of bile acids that are subject to extensive bacterial transformation in the intestinal tract. This could be of particular relevance as bile acids inhibit many bacteria and may modulate microbiota composition and also have the potential to affect major pathophysiological factors in IBS, including GI motility, secretion, and immune function (101,102). Indeed, in a recent large-scale study, increased colonic bile acid exposure was demonstrated in a subset of, predominantly non-constipated, IBS patients correlating with diarrhea and accelerated colonic transit (101).

An association between the intestinal microbiota and protein metabolism has been established by a significant correlation between the abundance of several post-infectious (PI)-IBS microbial markers and host amino-acid metabolism (10). Among harmful products of protein fermentation (100,103), hydrogen sulfide might be relevant for compromising intestinal health as it directly impairs epithelial metabolism (104) and acts as a gut–brain signaling molecule (105). Hydrogen sulfide can be converted to thiosulfate and further oxidized to tetrathionate during inflammation. The latter supports the growth of *Salmonella* (106) and other tetrathionate utilizing pathogens from the Gamma-Proteobacteria class (107), many of which have been associated with bowel symptoms of IBS patients (9,10,82).

FODMAPS, indigestible carbohydrates, can be fermented by intestinal microbes, resulting in increased gas production and an osmotic effect that can provoke IBS symptoms in a sensitive host (108), although similar changes may go unperceived in normal subjects (109). Carbohydrate utilizing GI bacteria such as *Dorea* spp., major gas producing bacteria in the human intestine (110), were reported to be significantly increased in abundance in IBS patients (9,11). Overproduction of gas, specifically hydrogen, has been associated with IBS symptoms (111), especially abdominal pain and flatulence. In patients with IBS-D, whose colon was found to be less able to accommodate the increased intestinal volume (112), overproduction of gas could be a trigger for increased colonic wall tension. In general, hydrogen produced in the gut is most efficiently removed by the methanogenic archaea (113) that appear to be depleted in IBS-D patients (9,10) and increased in IBS with constipation patients (reviewed in (114)) Although there is substantial evidence for implication of methane in IBS symptomology, most data are based upon methane breath testing. Hence, future studies, especially those addressing the diet-microbiota connection in IBS, should quantify methanogens to provide more direct evidence.

Alternative pathway for gas elimination is the conversion into acetate by acetogenic bacteria, *Blautia* spp. (115). *Blautia* spp. are highly prevalent and dominant intestinal bacteria, which have been found to be elevated in IBS patients (9,79). Intriguingly, the major fermentation end-product—acetate—has been reported as significantly increased in IBS patients (116), although this was not confirmed in another study (117). This may relate to the fact that most studies measure stool concentrations without taking the total stool volume into account, thereby potentially concealing overall increased acetate production as a result of dilution due to increased stool volume.

SCFAs produced upon the microbial fermentation of undigested carbohydrates and reduced levels thereof may also contribute to symptom generation in IBS. SCFAs have some documented health benefits, in particular butyrate, which is an important energy source for colonocytes, and among others inhibits inflammation and enhances barrier function (118). Moreover, intraluminal administration of butyrate into the distal colon has been shown to decrease visceral sensitivity in healthy humans (119). However, excessive intake of fermentable carbohydrates and high SCFA levels can also increase the osmotic load and thereby lead to diarrhea (120).

Data on SCFAs in IBS are limited and inconsistent, showing no alterations, increased or decreased levels of SCFAs relative to controls (61,116,117,121). Furthermore, altered SCFAs profiles might merely reflect diet rather than a feature of the condition. Given that the major products of FODMAPs fermentation are SCFAs, and that some major microbial SCFA-producers may be altered in IBS, the connection between FODMAPs, SCFAs, microbiota, and IBS symptoms calls for further elucidation.

Few studies have examined the impact of dietary interventions, in particular FODMAP restriction, on the microbiota in IBS patients. A low-FODMAP diet has been linked to reduced bifidobacterial counts (59), which seems a paradox given their potential health benefit. A recent study reported a reduction in total bac-

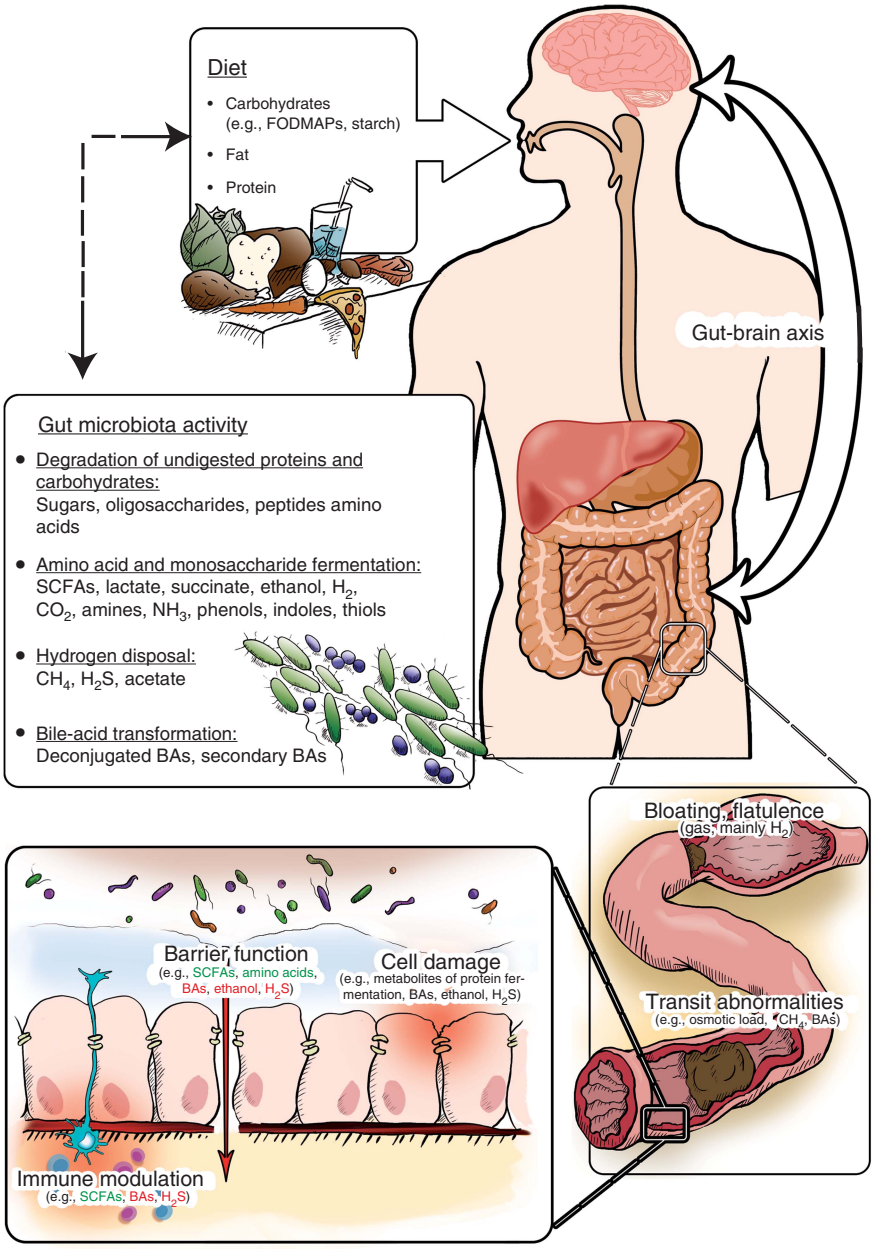


Figure 2. Non-exclusive listing of dietary metabolites that can be regulated by microbial activity and contribute to irritable bowel syndrome (IBS) symptoms. The impact of metabolites affecting different domains of intestinal health is depicted; however, it should be noted that the effects of some metabolites depend on their concentration or further conversion. Note: for immune modulation and barrier function, metabolites with positive and negative effects are colored in green and red, respectively. SCFAs, short-chain fatty acids; BAs, bile acids.

terial abundance following the introduction of a low-FODMAP diet, as well as a reduction in absolute counts but not relative abundance of specific bacterial groups, including bifidobacteria as compared with habitual diet (63). However, as compared with a diet high in FODMAP content, the low-FODMAP diet appeared to be associated with marked lower relative abundances of butyrate-producing bacteria and *A. muciniphila*, and a significantly higher abundance of *R. torques*, adding more paradoxical data on the effects of low-FODMAP diet (63).

Extensive analyses of microbiota composition, functionality, and fermentation products in relation to FODMAP restriction and symptom generation are currently, however, lacking.

The direct impact of diet-microbiota interactions on IBS symptoms has been demonstrated in an elegant study on the impact of a diet rich in fermentable substrates. The introduction of this flatulogenic diet significantly increased the gas volume and the number of gas evacuations, as well as abdominal symptoms and digestive discomfort in both healthy subjects and flatulent

© 2015 by the American College of Gastroenterology

patients (122). Yet, upon dietary challenge, patients' microbiota developed instability in composition, exhibiting variations in the *Bacteroidetes* and *Firmicutes* phyla and reduction in microbial diversity, whereas healthy subjects' microbiota remained stable. Hence, the extent of dysbiosis in IBS patients likely depends on their diet. In a small study among children with IBS, the introduction of a low fermentable substrate diet significantly decreased abdominal pain severity and frequency in some but not all children. Responders to this dietary intervention appeared to differ in intestinal microbiota composition at baseline from nonresponders, indicating that the efficacy of dietary interventions might be influenced by the patient's microbiota (123). To prove the validity, these pilot results will have to be replicated in larger studies.

CONCLUSION, RECOMMENDATIONS FOR FUTURE RESEARCH

Both experimental and observational studies provide good evidence to conclude that microbial alterations frequently observed in IBS patients potentially drive or perpetuate gastrointestinal symptoms. Yet, the strong interaction between distinct dietary patterns and the intestinal microbial communities is likely one of the explanations for the inconclusive findings among studies comparing the microbiota composition in IBS patients and healthy subjects.

More well-designed studies on the effect of food intake on symptom generation and the underlying mechanisms are warranted. The potential benefit of FODMAP restriction seems promising but requires further study and dietary guidance is essential given the complexity and risk of inadequate nutrient intake. In addition, more evidence on long-term outcome and well-performed challenge or re-introduction trials are warranted as important adaptions may occur. The incorporation of microbiota characterization in such studies would both provide insight into the impact of these dietary interventions on the microbiota, as well as offer the opportunity to examine whether the microbiota can predict response.

Given that IBS patients comprise a heterogeneous group both in the type of bowel symptoms and food components to which they might respond, and that the microbiota is extremely complex and variable between individuals, it is clear that defining the correlations between IBS, microbiota, and diet is not an easy task. However, high-throughput molecular methods enable reproducible and in-depth analysis of the intestinal microbiota, providing the necessary tool for testing these correlations. The major challenge remains to include the conscientious collection of dietary information in future studies in this field. Moreover, future studies should include multiple microbiota analysis at various time points in a large number of well-phenotyped patients to increase our insight in the contribution of intestinal microbiota perturbations to the diverse and fluctuating symptom pattern in IBS patients.

ACKNOWLEDGMENTS

We are thankful to Mayk Lucchesi for figure designing and formatting.

CONFLICT OF INTEREST

Guarantor of the article: John Penders, PhD.

Specific author contributions: The initial drafting of the manuscript was prepared by Mirjana Rajilić-Stojanović, Daisy M. Jonkers, Robin C. Spiller, Anne Salonen, Kurt Hanevik, Jonna Jalanka, Chaysavanh Manichanh, Elena Philippou, Gerard Clarke, and John Penders. Editing and draft revisions were done by Mirjana Rajilić-Stojanović, Daisy M. Jonkers, Anne Salonen, Kurt Hanevik, Jeroen Raes, Jonna Jalanka, Willem M. de Vos, Chaysavanh Manichanh, Natasa Golic, Paul Enck, Elena Philippou, Fuad A. Iraqi, Gerard Clarke, Robin C. Spiller, and John Penders. All authors have approved the final draft.

Financial Support: This work was supported by collaboration and network activities promoted under the frame of the international network GENIEUR (Genes in Irritable Bowel Syndrome Europe), which is currently funded by the COST program (BM1106, www.GENIEUR.eu).

Potential competing interests: Willem M. de Vos served as a member of the Scientific Advisory Boards of NIHS and Johnson & Johnson and has filed patent applications on IBS diagnostics. Paul Enck has received research funds from SymbioPharm, Germany, and has received consultancy fees and travel reimbursement from SymbioPharm, Germany, and Danone, France. Robin Spiller has received research funding from Lesaffre and Ironwood and free drug for clinical trial from Norgine and Falk Pharma. He has also acted on Advisory Boards for Almirall, Astellas, Yuhan Corporation, and Danone. Jeroen Raes served on advisory boards for Johnson & Johnson, GSK Vaccines, and 23andMe. The Alimentary Pharmabiotic Centre is a research center funded by Science Foundation Ireland (SFI; grant numbers SFI/12/RC/2273, 02/CE/B124 and 07/CE/B1368), through the Irish Government's National Development Plan. The Centre has conducted studies in collaboration with several companies including GSK, Pfizer, Wyeth, and Mead Johnson. Gerard Clarke is also supported by the Health Research Board (HRB) through a Health Research Award (grant no HRA_POR/2011/23) and by a NARSAD Young Investigator Grant from the Brain and Behavior Research Foundation (grant Number 20771). The content of this article was neither influenced nor constrained by this support.

REFERENCES

1. Lovell RM, Ford AC. Global prevalence of and risk factors for irritable bowel syndrome: a meta-analysis. Clin Gastroenterol Hepatol 2012;10:712–21.
2. Simrén M, Svedlund J, Posserud I *et al.* Health-related quality of life in patients attending a gastroenterology outpatient clinic: functional disorders versus organic diseases. Clin Gastroenterol Hepatol 2006;4:187–95.
3. Longstreth GF, Thompson WG, Chey WD *et al.* Functional bowel disorders. Gastroenterology 2006;130:1480–91.
4. Camilleri M, Lasch K, Zhou W. Irritable bowel syndrome: methods, mechanisms, and pathophysiology. The confluence of increased permeability, inflammation, and pain in irritable bowel syndrome. Am J Physiol Gastrointest Liver Physiol 2012;303:G775–G785.
5. Thabane M, Kottachchi DT, Marshall JK. Systematic review and meta-analysis: The incidence and prognosis of post-infectious irritable bowel syndrome. Aliment Pharmacol Ther 2007;26:535–44.
6. Villarreal AA, Aberger FJ, Benrud R *et al.* Use of broad-spectrum antibiotics and the development of irritable bowel syndrome. WMJ 2012;111:17–20.
7. Salonen A, de Vos WM, Palva A. Gastrointestinal microbiota in irritable bowel syndrome: present state and perspectives. Microbiology 2010;156:3205–15.

8. Jeffery IB, O'Toole PW, Öhman L *et al.* An irritable bowel syndrome subtype defined by species-specific alterations in faecal microbiota. Gut 2012;61:997–1006.
9. Rajilić-Stojanović M, Biagi E, Heilig HGHJ *et al.* Global and deep molecular analysis of microbiota signatures in fecal samples from patients with irritable bowel syndrome. Gastroenterology 2011;141:1792–801.
10. Jalanka-Tuovinen J, Salojärvi J, Salonen A *et al.* Faecal microbiota composition and host-microbe cross-talk following gastroenteritis and in postinfectious irritable bowel syndrome. Gut 2014;63:1737–45.
11. Saulnier DM, Riehle K, Mistretta T-A *et al.* Gastrointestinal microbiome signatures of pediatric patients with irritable bowel syndrome. Gastroenterology 2011;141:1782–91.
12. Funkhouser LJ, Bordenstein SR. Mom knows best: the universality of maternal microbial transmission. PLoS Biol 2013;11:e1001631.
13. Aagaard K, Ma J, Antony KM *et al.* The placenta harbors a unique microbiome. Sci Transl Med 2014;6:237ra65.
14. Dominguez-Bello MG, Costello EK, Contreras M *et al.* Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. Proc Natl Acad Sci USA 2010;107:11971–5.
15. Penders J, Thijs C, Vink C *et al.* Factors influencing the composition of the intestinal microbiota in early infancy. Pediatrics 2006;118:511–21.
16. Penders J, Gerhold K, Stobberingh EE *et al.* Establishment of the intestinal microbiota and its role for atopic dermatitis in early childhood. J Allergy Clin Immunol 2013;132:601–7.e8
17. Azad MB, Konya T, Maughan H *et al.* Gut microbiota of healthy Canadian infants: profiles by mode of delivery and infant diet at 4 months. CMAJ 2013;185:385–94.
18. Fallani M, Young D, Scott J *et al.* Intestinal microbiota of 6-week-old infants across Europe: geographic influence beyond delivery mode, breast-feeding, and antibiotics. J Pediatr Gastroenterol Nutr 2010;51:77–84
19. Marcobal A, Barboza M, Froehlich JW *et al.* Consumption of human milk oligosaccharides by gut-related microbes. J Agric Food Chem 2010;58:5334–40.
20. Jost T, Lacroix C, Braegger C *et al.* Assessment of bacterial diversity in breast milk using culture-dependent and culture-independent approaches. Br J Nutr 2013;110:1253–62.
21. Palmer C, Bik EM, Digiulio DB *et al.* Development of the human infant intestinal microbiota. PLoS Biol 2007;5:e177.
22. Koenig JE, Spor A, Scalfone N *et al.* Succession of microbial consortia in the developing infant gut microbiome. Proc Natl Acad Sci USA 2011;108(Suppl 1):4578–85.
23. Nylund L, Satokari R, Salminen S *et al.* Intestinal microbiota during early life—impact on health and disease. Proc Nutr Soc 2014;73:457–69.
24. Borre YE, O'Keeffe GW, Clarke G *et al.* Microbiota and neurodevelopmental windows: implications for brain disorders. Trends Mol Med 2014;20:509–18.
25. Rajilić-Stojanović M, Heilig HGHJ, Tims S *et al.* Long-term monitoring of the human intestinal microbiota composition. Environ Microbiol 2013;15:1146–59.
26. Yatsunenko T, Rey FE, Manary MJ *et al.* Human gut microbiome viewed across age and geography. Nature 2012;486:222–7.
27. Costello EK, Lauber CL, Hamady M *et al.* Bacterial community variation in human body habitats across space and time. Science 2009;326:1694–7.
28. Zoetendal EG, Akkermans AD, de Vos WM. Temperature gradient gel electrophoresis analysis of 16S rRNA from human fecal samples reveals stable and host-specific communities of active bacteria. Appl Environ Microbiol 1998;64:3854–9.
29. Ding T, Schloss PD. Dynamics and associations of microbial community types across the human body. Nature 2014;509:357–60.
30. Lin A, Bik EM, Costello EK *et al.* Distinct distal gut microbiome diversity and composition in healthy children from Bangladesh and the United States. PLoS ONE 2013;8:e53838.
31. Schnorr SL, Candela M, Rampelli S *et al.* Gut microbiome of the Hadza hunter-gatherers. Nat Commun 2014;5:3654.
32. De Filippo C, Cavalieri D, Di Paola M *et al.* Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. Proc Natl Acad Sci 2010;107:14691–6.
33. Salonen A, de Vos WM. Impact of diet on human intestinal microbiota and health. Annu Rev Food Sci Technol 2014;5:239–62.
34. Wu GD, Chen J, Hoffmann C *et al.* Linking long-term dietary patterns with gut microbial enterotypes. Science 2011;334:105–8.
35. Lappi J, Salojärvi J, Kolehmainen M *et al.* Intake of whole-grain and fiber-rich rye bread versus refined wheat bread does not differentiate intestinal microbiota composition in Finnish adults with metabolic syndrome. J Nutr 2013;143:648–55.
36. Cotillard A, Kennedy SP, Kong LC *et al.* Dietary intervention impact on gut microbial gene richness. Nature 2013;500:585–8.
37. David LA, Maurice CF, Carmody RN *et al.* Diet rapidly and reproducibly alters the human gut microbiome. Nature 2014;505:559–63.
38. Salonen A, Lahti L, Salojarvi J *et al.* Impact of diet and individual variation on intestinal microbiota composition and fermentation products in obese men. ISME J 2014;8:2218–30.
39. Walker AW, Ince J, Duncan SH *et al.* Dominant and diet-responsive groups of bacteria within the human colonic microbiota. ISME J 2011;5:220–30.
40. Monsbakken KW, Vandvik PO, Farup PG. Perceived food intolerance in subjects with irritable bowel syndrome—etiology, prevalence and consequences. Eur J Clin Nutr 2006;60:667–72.
41. Simrén M, Mansson A, Langkilde AM *et al.* Food-related gastrointestinal symptoms in the irritable bowel syndrome. Digestion 2001;63:108–15.
42. Hayes PA, Fraher MH, Quigley EM. Irritable bowel syndrome: the role of food in pathogenesis and management. Gastroenterol Hepatol 2014;10:164–74.
43. Böhn L, Störsrud S, Simrén M. Nutrient intake in patients with irritable bowel syndrome compared with the general population. Neurogastroenterol Motil 2013;25:23–30
44. Saito YA, Locke GR, Weaver AL *et al.* Diet and functional gastrointestinal disorders: a population-based case control study. Am J Gastroenterol 2005;100:2743–8.
45. Williams EA, Nai X, Corfe BM. Dietary intakes in people with irritable bowel syndrome. BMC Gastroenterol 2011;11:9.
46. Ragnarsson G, Bodemar G. Pain is temporally related to eating but not to defaecation in the irritable bowel syndrome (IBS). Patients' description of diarrhea, constipation and symptom variation during a prospective 6-week study. Eur J Gastroenterol Hepatol 1998;10:415–21.
47. Simrén M, Agerforz P, Bjornsson ES *et al.* Nutrient-dependent enhancement of rectal sensitivity in irritable bowel syndrome (IBS). Neurogastroenterol Motil 2007;19:20–29.
48. Ludidi S, Conchillo JM, Keszthelyi D *et al.* Does meal ingestion enhance sensitivity of visceroperception assessment in irritable bowel syndrome? Neurogastroenterol Motil 2012;24:47–53e3
49. Böhn L, Storsrud S, Tornblom H *et al.* Self-reported food-related gastrointestinal symptoms in IBS are common and associated with more severe symptoms and reduced quality of life. Am J Gastroenterol 2013;108:634–41.
50. Niec AM, Frankum B, Talley NJ. Are adverse food reactions linked to irritable bowel syndrome? Am J Gastroenterol 1998;93:2184–90.
51. Atkins FM. A critical evaluation of clinical trials in adverse reactions to foods in adults. J Allergy Clin Immunol 1986;78:174–82.
52. Young E, Stoneham MD, Petruckevitch A *et al.* A population study of food intolerance. Lancet 1994;343:1127–30.
53. Carroccio A, Brusca I, Mansueto P *et al.* Fecal assays detect hypersensitivity to cow's milk protein and gluten in adults with irritable bowel syndrome. Clin Gastroenterol Hepatol 2011;9:965–71.e3
54. Feinle-Bisset C, Azpiroz F. Dietary lipids and functional gastrointestinal disorders. Am J Gastroenterol 2013;108:737–47.
55. Yang J, Fox M, Cong Y *et al.* Lactose intolerance in irritable bowel syndrome patients with diarrhoea: the roles of anxiety, activation of the innate mucosal immune system and visceral sensitivity. Aliment Pharmacol Ther 2014;39:302–11.
56. McKenzie YA, Alder A, Anderson W *et al.* British Dietetic Association evidence-based guidelines for the dietary management of irritable bowel syndrome in adults. J Hum Nutr Diet 2012;25:260–74.
57. Shepherd SJ, Gibson PR. Fructose malabsorption and symptoms of irritable bowel syndrome: guidelines for effective dietary management. J Am Diet Assoc 2006;106:1631–9.
58. de Roest RH, Dobbs BR, Chapman BA *et al.* The low FODMAP diet improves gastrointestinal symptoms in patients with irritable bowel syndrome: a prospective study. Int J Clin Pract 2013;67:895–903
59. Staudacher HM, Lomer MC, Anderson JL *et al.* Fermentable carbohydrate restriction reduces luminal bifidobacteria and gastrointestinal symptoms in patients with irritable bowel syndrome. J Nutr 2012;142:1510–8.
60. Ong DK, Mitchell SB, Barrett JS *et al.* Manipulation of dietary short chain carbohydrates alters the pattern of gas production and genesis of symptoms in irritable bowel syndrome. J Gastroenterol Hepatol 2010;25:1366–73.

© 2015 by the American College of Gastroenterology

61. Halmos EP, Power VA, Shepherd SJ *et al.* A diet low in FODMAPs reduces symptoms of irritable bowel syndrome. Gastroenterology 2014;146:67–75; e5.
62. Shepherd SJ, Parker FC, Muir JG *et al.* Dietary triggers of abdominal symptoms in patients with irritable bowel syndrome: randomized placebo-controlled evidence. Clin Gastroenterol Hepatol 2008;6:765–71.
63. Halmos EP, Christophersen CT, Bird AR *et al.* Diets that differ in their FODMAP content alter the colonic luminal microenvironment. Gut 2014;64:93–100.
64. Ligaarden SC, Lydersen S, Farup PG. IgG and IgG4 antibodies in subjects with irritable bowel syndrome: a case control study in the general population. BMC Gastroenterol 2012;12:166.
65. Spiller RC, Jenkins D, Thornley JP *et al.* Increased rectal mucosal enteroendocrine cells, T lymphocytes, and increased gut permeability following acute Campylobacter enteritis and in post-dysenteric irritable bowel syndrome. Gut 2000;47:804–11.
66. Walker MM, Warwick A, Ung C *et al.* The role of eosinophils and mast cells in intestinal functional disease. Curr Gastroenterol Rep 2011;13:323–30.
67. Matricon J, Meleine M, Gelot A *et al.* Review article: associations between immune activation, intestinal permeability and the irritable bowel syndrome. Aliment Pharmacol Ther 2012;36:1009–31.
68. Ford AC, Chey WD, Talley NJ *et al.* Yield of diagnostic tests for celiac disease in individuals with symptoms suggestive of irritable bowel syndrome: systematic review and meta-analysis. Arch Intern Med 2009;169:651–8.
69. Wahnschaffe U, Schulzke JD, Zeitz M *et al.* Predictors of clinical response to gluten-free diet in patients diagnosed with diarrhea-predominant irritable bowel syndrome. Clin Gastroenterol Hepatol 2007;5:844–50l;quiz 769.
70. Biesiekierski JR, Newnham ED, Irving PM *et al.* Gluten causes gastrointestinal symptoms in subjects without celiac disease: a double-blind randomized placebo-controlled trial. Am J Gastroenterol 2011;106:508–14;quiz 515.
71. Biesiekierski JR, Peters SL, Newnham ED *et al.* No effects of gluten in patients with self-reported non-celiac gluten sensitivity after dietary reduction of fermentable, poorly absorbed, short-chain carbohydrates. Gastroenterology 2013;145:320–8;e1-3.
72. Carroccio A, Mansueto P, D'Alcamo A *et al.* Non-celiac wheat sensitivity as an allergic condition: personal experience and narrative review. Am J Gastroenterol 2013;108:1845–52;quiz 1853.
73. Ford AC, Quigley EM, Lacy BE *et al.* Efficacy of prebiotics, probiotics, and synbiotics in irritable bowel syndrome and chronic idiopathic constipation: systematic review and meta-analysis. Am J Gastroenterol 2014;109:1547–61.
74. Moayyedi P, Quigley EM, Lacy BE *et al.* The effect of fiber supplementation on irritable bowel syndrome: a systematic review and meta-analysis. Am J Gastroenterol 2014;109:1367–74.
75. Kassinen A, Krogius-Kurikka L, Mäkivuokko H *et al.* The fecal microbiota of irritable bowel syndrome patients differs significantly from that of healthy subjects. Gastroenterology 2007;133:24–33.
76. Simrén M, Barbara G, Flint HJ *et al.* Intestinal microbiota in functional bowel disorders: a Rome foundation report. Gut 2013;62:159–76.
77. Krogius-Kurikka L, Lyra A, Malinen E *et al.* Microbial community analysis reveals high level phylogenetic alterations in the overall gastrointestinal microbiota of diarrhoea-predominant irritable bowel syndrome sufferers. BMC Gastroenterol 2009;9:95.
78. Kerckhoffs APM, Samsom M, van der Rest ME *et al.* Lower *Bifidobacteria* counts in both duodenal mucosa-associated and fecal microbiota in irritable bowel syndrome patients. World J Gastroenterol 2009;15:2887–92.
79. Malinen E, Rinttilä T, Kajander K *et al.* Analysis of the fecal microbiota of irritable bowel syndrome patients and healthy controls with real-time PCR. Am J Gastroenterol 2005;100:373–82.
80. Balsari A, Ceccarelli A, Dubini F *et al.* The fecal microbial population in the irritable bowel syndrome. Microbiologica 1982;5:185–94.
81. Si JM, Yu YC, Fan YJ *et al.* Intestinal microecology and quality of life in irritable bowel syndrome patients. World J Gastroenterol 2004;10:1802–5.
82. Carroll IM, Shen XJ, Keku TO *et al.* Characterization of the fecal microbiota in patients with diarrhea predominant irritable bowel syndrome. Gastroenterology 2008;134:A-681.
83. Durbán A, Abellán JJ, Jiménez-Hernández N *et al.* Instability of the faecal microbiota in diarrhoea-predominant irritable bowel syndrome. FEMS Microbiol Ecol 2013;86:581–9.
84. Mättö J, Maunuksela L, Kajander K *et al.* Composition and temporal stability of gastrointestinal microbiota in irritable bowel syndrome—a longitudinal study in IBS and control subjects. FEMS Immunol Med Microbiol 2005;43:213–22.
85. Maukonen J, Satokari R, Mättö J *et al.* Prevalence and temporal stability of selected clostridial groups in irritable bowel syndrome in relation to predominant faecal bacteria. J Med Microbiol 2006;55:625–33.
86. Crouzet L, Gaultier E, Del'Homme C *et al.* The hypersensitivity to colonic distension of IBS patients can be transferred to rats through their fecal microbiota. Neurogastroenterol Motil 2013;25:e272–e282.
87. De Palma G. Transfer of anxiety and gut dysfunction from IBS patients to gnotobiotic mice through microibota transplatation. CDDW 2014; 2014:A183.
88. Buhner S, Li Q, Vignali S *et al.* Activation of human enteric neurons by supernatants of colonic biopsy specimens from patients with irritable bowel syndrome. Gastroenterology 2009;137:1425–34.
89. Gecse K, Roka R, Ferrier L *et al.* Increased fecal serine-protease activity in diarrheic IBS patients: a colonic lumenal factor impairing colonic permeability and sensitivity. Gut 2008;57:591–9.
90. Tooth D, Garsed K, Singh G *et al.* Characterisation of faecal protease activity in irritable bowel syndrome with diarrhoea: origin and effect of gut transit. Gut 2014;63:753–60.
91. Macfarlane GT, Allison C, Gibson SA *et al.* Contribution of the microflora to proteolysis in the human large intestine. J Appl Bacteriol 1988;64:37–46.
92. Tooth D, Garsed K, Singh G *et al.* Characterisation of faecal protease activity in irritable bowel syndrome with diarrhoea: origin and effect of gut transit. Gut 2014;63:753–60.
93. Ivanov D, Emonet C, Foata F *et al.* A serpin from the gut bacterium *Bifidobacterium longum* inhibits eukaryotic elastase-like serine proteases. J Biol Chem 2006;281:17246–52.
94. Lee KN, Lee OY. Intestinal microbiota in pathophysiology and management of irritable bowel syndrome. World J Gastroenterol 2014;20:8886–97.
95. Mayer EA, Savidge T, Shulman RJ. Brain-gut microbiome interactions and functional bowel disorders. Gastroenterology 2014;146:1500–12.
96. Rajilić-Stojanović M, Smidt H, de Vos WM. Diversity of the human gastrointestinal tract microbiota revisited. Environ Microbiol 2007;9:2125–36.
97. Malinen E, Krogius-Kurikka L, Lyra A *et al.* Association of symptoms with gastrointestinal microbiota in irritable bowel syndrome. World J Gastroenterol 2010;16:4532–40.
98. Lyra A, Krogius-Kurikka L, Nikkila J *et al.* Effect of a multispecies probiotic supplement on quantity of irritable bowel syndrome-related intestinal microbial phylotypes. BMC Gastroenterol 2009;10:110.
99. Rajilić-Stojanović M, Guarner F, Shanahan F *et al.* Phylogenetic analysis of dysbiosis in ulcerative colitis during remission. Inflamm Bowel Dis 2013;19:481–8.
100. Smith EA, Macfarlane GT. Dissimilatory amino acid metabolism in human colonic bacteria. Anaerobe 1997;3:327–37.
101. Bajor A, Törnblom H, Rudling M *et al.* Increased colonic bile acid exposure: a relevant factor for symptoms and treatment in IBS. Gut 2014;64:84–92.
102. Hylemon PB, Zhou H, Pandak WM *et al.* Bile acids as regulatory molecules. J Lipid Res 2009;50:1509–20.
103. Rajilić-Stojanović M. Function of the microbiota. Best Pract Res Clin Gastroenterol 2013;27:5–16.
104. Jørgensen J, Mortensen P. Hydrogen sulfide and colonic epithelial metabolism. Dig Dis Sci 2001;46:1722–32.
105. Schicho R, Krueger D, Zeller F *et al.* Hydrogen sulfide is a novel prosecretory neuromodulator in the guinea-pig and human colon. Gastroenterology 2006;131:1542.
106. Thiennimitr P, Winter SE, Winter MG *et al.* Intestinal inflammation allows *Salmonella* to use ethanolamine to compete with the microbiota. Proc Natl Acad Sci USA 2011;108:17480–5.
107. Weissfeld AS, Sonnenwirth AC. Rapid isolation of *Yersinia* spp. from feces. J Clin Microbiol 1982;15:508–10.
108. Staudacher HM, Irving PM, Lomer MC *et al.* Mechanisms and efficacy of dietary FODMAP restriction in IBS. Nat Rev Gastroenterol Hepatol 2014;11:256–66.
109. Murray K, Wilkinson-Smith V, Hoad C *et al.* Differential effects of FODMAPs (fermentable oligo-, di-, mono-saccharides and polyols) on small and large intestinal contents in healthy subjects shown by MRI. Am J Gastroenterol 2014;109:110–9.
110. Taras D, Simmering R, Collins MD *et al.* Reclassification of *Eubacterium formicigenerans* Holdeman and Moore 1974 as *Dorea formicigenerans* gen. nov., comb. nov., and description of *Dorea longicatena* sp. nov., isolated from human faeces. Int J Syst Evol Microbiol 2002;52:423–8.
111. King TS, Elia M, Hunter JO. Abnormal colonic fermentation in irritable bowel syndrome. Lancet 1998;352:1187–9.

112. Pritchard SE, Marciani L, Garsed KC *et al.* Fasting and postprandial volumes of the undisturbed colon: normal values and changes in diarrhea-predominant irritable bowel syndrome measured using serial MRI. Neurogastroenterol Motil 2014;26:124–30.
113. Pimentel M, Gunsalus RP, Rao SSC *et al.* Methanogens in human health and disease. Am J Gastroenterol Suppl 2012;1:28–33.
114. Triantafyllou K, Chang C, Pimentel M. Methanogens, methane and gastrointestinal motility. J Neurogastroenterol Motil 2014;20:31–40.
115. Liu C, Finegold SM, Song Y *et al.* Reclassification of *Clostridium coccoides, Ruminococcus hansenii, Ruminococcus hydrogenotrophicus, Ruminococcus luti, Ruminococcus productus* and *Ruminococcus schinkii* as *Blautia coccoides* gen. nov., comb. nov., *Blautia hansenii* comb. nov., *Blautia hydrogenotrophica* comb. nov., *Blautia luti* comb. nov., *Blautia producta* comb. nov., *Blautia schinkii* comb. nov. and description of *Blautia wexlerae* sp. nov., isolated from human faeces. Int J Syst Evol Microbiol 2008;58:1896–902.
116. Tana C, Umesaki Y, Imaoka A *et al.* Altered profiles of intestinal microbiota and organic acids may be the origin of symptoms in irritable bowel syndrome. Neurogastroenterol Motil 2010;22:512–9.
117. Treem WR, Ahsan N, Kastoff G *et al.* Fecal short-chain fatty acids in patients with diarrhea-predominant irritable bowel syndrome: *in vitro* studies of carbohydrate fermentation. J Pediatr Gastroenterol Nutr 1996;23:280–6.
118. Hamer HM, Jonkers D, Venema K *et al.* Review article: the role of butyrate on colonic function. Aliment Pharmacol Ther 2008;27:104–19.
119. Vanhoutvin SALW, Troost FJ, Kilkens TOC *et al.* The effects of butyrate enemas on visceral perception in healthy volunteers. Neurogastroenterol Motil 2009;21:952–e76.
120. Fritz E, Hammer HF, Lipp RW *et al.* Effects of lactulose and polyethylene glycol on colonic transit. Aliment Pharmacol Ther 2005;21:259–68.
121. Mortensen PB, Andersen JR, Arffmann S *et al.* Short-chain fatty acids and the irritable bowel syndrome: the effect of wheat bran. Scand J Gastroenterol 1987;22:185–92.
122. Manichanh C, Eck A, Varela E *et al.* Anal gas evacuation and colonic microbiota in patients with flatulence: effect of diet. Gut 2014;63:401–8.
123. Chumpitazi BP, Hollister EB, Oezguen N *et al.* Gut microbiota influences low fermentable substrate diet efficacy in children with irritable bowel syndrome. Gut Microbes 2014;5:165–75.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/3.0/

© 2015 by the American College of Gastroenterology

Altered Colonic Bacterial Fermentation as a Potential Pathophysiological Factor in Irritable Bowel Syndrome

Tamar Ringel-Kulka, MD, MPH[1,6], Chang Hwan Choi, MD, PhD[2,3,6], Daniel Temas, BS[2], Ari Kim, MD, PhD[1,4], Daniele M. Maier, MS, C-PA[2], Karen Scott, PhD[5], Joseph A. Galanko, PhD[2] and Yehuda Ringel, MD[2]

OBJECTIVES: **Dysbiosis leading to abnormal intestinal fermentation has been suggested as a possible etiological mechanism in irritable bowel syndrome (IBS). We aimed to investigate the location and magnitude of altered intestinal bacterial fermentation in IBS and its clinical subtypes.**

METHODS: **IBS patients who satisfied the Rome III criteria (114) and 33 healthy controls (HC) were investigated. Intestinal fermentation was assessed using two surrogate measures: intestinal intraluminal pH and fecal short-chain fatty acids (SCFAs). Intraluminal pH and intestinal transit times were measured in the small and large bowel using a wireless motility capsule (SmartPill) in 47 IBS and 10 HC. Fecal SCFAs including acetate, propionate, butyrate, and lactate were analyzed by capillary gas chromatography in all enrolled subjects. Correlations between intestinal pH, fecal SCFAs, intestinal transit time, and IBS symptom scores were analyzed.**

RESULTS: **Colonic intraluminal pH levels were significantly lower in IBS patients compared with HC (total colonic pH, 6.8 for IBS vs. 7.3 for HC, *P*=0.042). There were no differences in total and segmental pH levels in the small bowel between IBS patients and HC (6.8 vs. 6.8, *P*=not significant). The intraluminal colonic pH differences were consistent in all IBS subtypes. Total SCFA level was significantly lower in C-IBS patients than in D-IBS and M-IBS patients and HC. The total SCFA level in all IBS subjects was similar with that of HC. Colonic pH levels correlated positively with colon transit time (CTT) and IBS symptoms severity. Total fecal SCFAs levels correlated negatively with CTT and positively with stool frequency.**

CONCLUSIONS: **Colonic intraluminal pH is decreased, suggesting higher colonic fermentation, in IBS patients compared with HC. Fecal SCFAs are not a sensitive marker to estimate intraluminal bacterial fermentation.**

Am J Gastroenterol 2015; 110:1339–1346; doi:10.1038/ajg.2015.220; published online 25 August 2015

INTRODUCTION

Irritable bowel syndrome (IBS) is characterized by chronic or recurring abdominal pain or discomfort associated with altered bowel habits and is one of the most common gastrointestinal (GI) disorders (1,2). IBS is considered a multifactorial disorder associated with visceral hypersensitivity, altered gut motility, and dysfunction of the brain–gut axis and immune system. However, the pathophysiology of the disorder is still not completely understood (2–5). Recently, it has been suggested that alterations in the gut microbiota, leading to abnormal intestinal fermentation, may be a possible etiological mechanism (6–9).

Intestinal fermentation by gut microbiota is a central physiological process by which polymers, including carbohydrates, are biotransformed into end-products, mainly short-chain fatty acids (SCFAs) and gases. These intraluminal fermentation processes have an important role in supplying nutrients and energy to the host (10). However, alterations in intestinal fermentation may lead to certain physiological abnormalities that are often

[1]Gillings School of Global Public Health, University of North Carolina, Chapel Hill, North Carolina, USA; [2]Division of Gastroenterology and Hepatology, Center for Gastrointestinal Biology and Disease, University of North Carolina, Chapel Hill, North Carolina, USA; [3]Department of Internal Medicine, Chung-Ang University College of Medicine, Seoul, Korea; [4]Department of Obstetrics and Gynecology, Institute of Wonkwang Medical Science, Wonkwang University College of Medicine, Iksan, Korea; [5]Rowett Institute of Nutrition and Health, University of Aberdeen, Aberdeen, UK; [6]The first two authors contributed equally to this work.
Correspondence: Yehuda Ringel, MD, Division of Gastroenterology and Hepatology, Center for Gastrointestinal Biology and Disease, University of North Carolina, 4107 BioInformatics Bldg, CB# 7080, 130 Mason Farm RD, Chapel Hill, North Carolina 27599-7080, USA. E-mail: yehuda_ringel@med.unc.edu
Received 30 December 2014; accepted 20 June 2015

© 2015 by the American College of Gastroenterology

observed in IBS such as intraluminal excessive gas production and altered motility. These alterations in intestinal physiology can be an important factor in provoking or exacerbating IBS symptoms (11,12). In addition, interventional clinical studies targeting the intestinal microbiota with antibiotics or probiotics demonstrated beneficial effects in some patients with IBS (13,14). Specifically, in regard to intestinal fermentation, recent studies have shown that a diet low in FODMAPs (fermentable oligosaccharides, disaccharides, monosaccharides, and polyols) can offer considerable symptom relief in patients with IBS (15,16). However, the role of altered intestinal fermentation by intestinal bacteria in the pathogenesis of IBS has not been adequately investigated in this condition.

We hypothesized that bacterial fermentation is altered in certain segments of the GI tract in IBS and that these alterations are associated with bowel functions and symptoms. The aims of the current study were to investigate the location and magnitude of altered intestinal bacterial fermentation in IBS and its clinical subtypes and to examine their relation to bowel characteristics and GI symptoms. Intestinal fermentation was assessed by measuring intraluminal pH levels (by a wireless motility capsule (SmartPill; Given Imaging, Yoqneam, Israel)) and fecal SCFAs concentration as surrogate markers in a cohort of well-characterized patients with IBS and healthy controls (HC).

METHODS

Study population

All subjects were recruited from the University of North Carolina (UNC) healthcare outpatient clinics and from the Chapel Hill general population by advertising. Inclusion criteria included age of 18 years or older, any gender, race, or ethnicity. HCs had no chronic or recurring GI symptoms. IBS patients met the Rome III criteria for IBS and had mild-to-moderate symptom severity with the IBS-symptom severity scale (IBS-SSS) score of 175–300. Additional inclusion criteria included normal physical examination and laboratory tests and a normal colonoscopy within the last 3 years, in patients aged 50 years or older. Subjects with a history of GI tract surgery (other than appendectomy or cholecystectomy), a history of inflammatory bowel diseases, celiac disease, lactose malabsorption, or any other diagnosis that could explain their chronic or recurrent bowel symptoms were excluded from the study. In addition, participants were excluded if they had a history of treatment with antibiotics, anti-inflammatory agents—including aspirin, non-steroidal anti-inflammatory drugs or steroids, or intentional consumption of probiotics in the past 6 weeks. All subjects were evaluated by a physician to exclude an alternative diagnosis to IBS. The study was approved by the UNC internal review board, and all subjects provided written consent before participation and were compensated for participating in the study.

Clinical symptom assessments in IBS

Abdominal pain/discomfort, altered bowel habit (loose/watery or hard/lumpy stools), and bloating were assessed using a 10-cm visual analogue scale (maximum score, 10) (17). IBS-SSS (18) and IBS-specific quality-of-life (IBS-QOL) (19) were also assessed. Stool frequency was measured by the number of bowel movements per day and stool consistency determined using the Bristol Stool Scale (20).

Assessment for intraluminal fermentation and intestinal transit

Intestinal fermentation was assessed using two surrogate markers: intestinal intraluminal pH and fecal SCFAs. Intraluminal pH levels were measured using a wireless motility capsule (SmartPill). The wireless motility capsule is an FDA approved device that is clinically used to assess intestinal transit. After being swallowed, the capsule constantly measures intraluminal pH, pressure, and temperature as it travels through the GI tract and transmits this information wirelessly to a portable receiver worn by the examinee (21). Using these parameters and a specific software, the (SmartPill) system generates data on gastric emptying time, small bowel transit time (SBTT), colon transit time (CTT), and whole-gut transit time (WGTT) (21,22). For the purpose of this study, we used the intraluminal pH data as surrogate markers for intestinal fermentation. To enable further accuracy in assessment of localization of intestinal fermentation, the small and large bowel pH measurements were both divided into four quartiles (Q1–4) based on the total time in each (small and large bowel) segment. SBTT, CTT, and WGTT were also assessed as previously described (21,22).

Analysis of SCFA

Fresh stool samples were collected from all subjects on site or at home in the morning of the study visit. Each fecal sample was immediately transferred to the laboratory where it was homogenized and stored at −80 °C until analyzed. Fecal acetate, propionate, butyrate, lactate, and total SCFAs were quantified in fecal samples in duplicate using capillary gas chromatography with 2-ethyl butyric acid as an internal standard, as described previously (10,23–25).

Statistical analysis

We used Student's *t*-test when there were two groups being compared and analysis of variance when there were more than two groups being compared for continuous variables. The χ^2 test was used for discrete variables. For *post hoc* analysis, Tukey's honest significance test was used. Baseline characteristics were evaluated by Student's *t*-test or Pearson's χ^2 test. Intestinal pH, fecal SCFA levels, and GI transit times were compared between groups by Student's *t*-test or analysis of variance. Correlations between colonic pH, fecal SCFAs, CTT, stool frequency and consistency, IBS-SSS, and IBS-QOL were performed using Spearman's correlation analysis for all study subjects including IBS and HC. All *P*-values were two-tailed, and P-values<0.05 were considered statistically significant. Statistical analyses were performed using SPSS version 18.0 for Windows (SPSS, Chicago, IL, USA). Unless otherwise stated, data are presented as mean±s.d.

Table 1. Baseline characteristics of the study groups

	IBS	HC	*P* value
Number of subjects	114	33	
Age (years), mean	35.4±11.3	33.9±13.0	0.534
Sex (male:female)	16:98	4:29	0.778
Body mass index (kg/m^2), mean	27.1±5.5	27.2±9.2	0.929
Race			0.054
Caucasian	84 (73.7%)	20 (60.6%)	
Black	24 (21.1%)	7 (21.2%)	
Others	6 (5.3%)	6 (18.2%)	
IBS subtype (number)			
Constipation-predominant	26 (22.8%)		
Diarrhea-predominant	42 (36.8%)		
Mixed	32 (28.1%)		
Unspecified	14 (12.3%)		

HC, healthy control; IBS, irritable bowel syndrome.
The data are presented as mean±s.d.

Table 2. Small and large bowel intraluminal pH levels

	IBS (*n*=47)	HC (*n*=10)	*P* value
Small bowel pH			
Q1 small bowel pH	6.1±0.7	6.2±1.0	0.885
Q2 small bowel pH	6.7±0.5	6.7±0.4	0.840
Q3 small bowel pH	7.1±0.4	7.1±0.4	0.994
Q4 small bowel pH	7.3±0.4	7.3±0.5	0.878
Total small bowel pH	6.8±0.4	6.8±0.4	0.937
Colonic pH			
Q1 colonic pH	6.4±0.6	6.8±0.4	0.011
Q2 colonic pH	6.8±0.7	7.1±0.5	0.173
Q3 colonic pH	7.0±0.8	7.5±0.6	0.081
Q4 colonic pH	7.2±0.8	7.8±0.8	0.046
Total colonic pH	6.8±0.7	7.3±0.5	0.042

HC, healthy control; IBS, irritable bowel syndrome; Q, quartile.
The data are presented as mean±s.d.

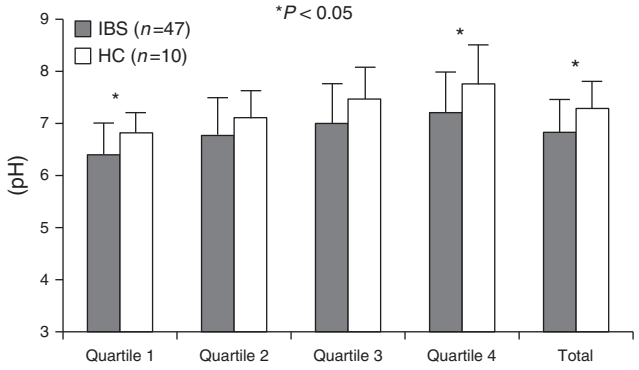


Figure 1. Colonic pH levels. Quartile 1 and 4 and mean total colonic pH levels were significantly lower in irritable bowel syndrome (IBS; *n*=47) compared with healthy control (HC; *n*=10). The pH in the proximal colon (quartile 1) was the lowest in both groups. **P*<0.05.

RESULTS

Study population

We investigated a total of 147 subjects (114 IBS and 33 HC). The study IBS population included patients with constipation-predominant IBS (C-IBS, *n*=26), diarrhea-predominant IBS (D-IBS, *n*=42), mixed type IBS (M-IBS, *n*=32), and unspecified type IBS (*n*=14). The racial distribution was as follows: Caucasian 70.7%, Black 21.1%, and other 8.2%. There were no significant differences in age, sex, body mass index, and race between the study groups (**Table 1**).

Intestinal intraluminal pH

Intestinal intraluminal pH data were collected from 47 patients with IBS and 10 HCs. Mean total colonic pH levels were significantly lower in the IBS group compared with HC (6.8 vs. 7.3, *P*=0.042). The significant differences were also observed in Q1 (6.4 vs. 6.8, *P*=0.011) and Q4 (7.2 vs. 7.8, *P*=0.046). The colonic pH levels at the Q1 were the lowest among the four quartiles in both groups (**Table 2** and **Figure 1**). The lower intraluminal colonic pH levels compared with HC were consistent across all IBS subtypes (mean total colonic pH: 7.0 for C-IBS, 7.0 for D-IBS, 6.8 for M-IBS, 6.4 for unspecified type IBS, and 7.3 for HC, respectively; **Table 3**).

There were no intraluminal pH differences in the four quartiles and total small bowel between IBS and HC (mean total small bowel pH: 6.8 for IBS vs. 6.8 for HC, *P*=0.937; **Table 2**). Also, there were no significant intraluminal small bowel pH differences between the groups based on IBS subtypes.

In subgroup analyses, small and large bowel pH levels were not different according to gender, race, and age (≤45 vs. >45).

Fecal SCFAs

Fecal SCFAs were measured in all enrolled subjects (114 IBS and 33 HC). Mean acetate, propionate, butyrate, lactate, and total fecal SCFAs levels were not significantly different between all IBS and HC (total SCFAs level: 92.1 mM vs. 92.0 mM, *P*=0.996, respectively). Acetate had the highest levels followed by propionate and butyrate in both IBS and HC. Lactate was not detected, except for two cases (6.4 mM in a M-IBS subject and 8.1 mM in a D-IBS subject). The SCFAs profiles were similar across the different IBS subtypes (**Table 4** and **Figure 2**). However, acetate, propionate, butyrate, and total SCFAs levels were lower in C-IBS than in all the other subtypes and were significantly lower compared with D-IBS. The propionate level in C-IBS was significantly lower than in any other group.

© 2015 by the American College of Gastroenterology

Table 3. Large bowel intraluminal pH levels according to IBS subtype

	C-IBS (*n*=13)	D-IBS (*n*=14)	M-IBS (*n*=12)	U-IBS (*n*=8)	HC (*n*=10)
Q1 colonic pH	6.4±0.5	6.7±0.7	6.4±0.5	5.9±0.3[a]	6.8±0.4
Q2 colonic pH	7.0±0.9	6.9±0.8	6.5±0.6	6.5±0.4	7.1±0.5
Q3 colonic pH	7.1±0.8	7.0±0.6	7.1±0.9	6.7±0.8	7.5±0.6
Q4 colonic pH	7.3±0.8	7.3±0.7	7.2±1.0	7.0±0.7	7.8±0.8
Total colonic pH	7.0±0.7	7.0±0.6	6.8±0.7	6.4±0.6[b]	7.3±0.5

C-IBS, constipation-predominant IBS; D-IBS, diarrhea-predominant IBS; HC, healthy control; IBS, irritable bowel syndrome; M-IBS, mixed type IBS; Q, quartile; U-IBS, unspecified type IBS.
The data are presented as mean±s.d.
[a]*P*<0.05 compared with Q1 colonic pHs of D-IBS and HC.
[b]*P*<0.05 compared with total colonic pH of HC.

Table 4. Fecal SCFAs levels

	C-IBS (*n*=26)	D-IBS (*n*=42)	M-IBS (*n*=32)	U-IBS (*n*=14)	IBS total (*n*=114)	HC (*n*=33)	*P* value[a]
Acetate (mM)	47.8±23.4[b]	61.3±23.2	59.2±21.2	63.9±40.9	58.0±25.8	56.6±18.5	0.777
Propionate (mM)	13.6±5.8[c]	20.1±8.9	18.4±10.2	20.4±14.5	18.2±9.8	20.0±8.0	0.331
Butyrate (mM)	11.6±9.5[d]	16.4±8.9	15.8±10.5	21.5±22.8	15.8±12.1	15.4±8.9	0.875
Total SCFAs (mM)	73.0±35.7[e]	98.1±36.7	93.7±37.0	105.9±77.0	92.1±44.1	92.0±33.3	0.996

C-IBS, constipation-predominant IBS; D-IBS, diarrhea-predominant IBS; HC, healthy control; IBS, irritable bowel syndrome; M-IBS, mixed type IBS; SCFA, short-chain fatty acid; U-IBS, unspecified type IBS.
The data are presented as mean±s.d.
[a]*P*-value for comparison between total IBS and HC.
[b]*P*<0.05 compared with the acetate level of D-IBS.
[c]*P*<0.05 compared with propionate levels of all the other groups.
[d]*P*<0.05 compared with the butyrate level of D-IBS.
[e]*P*<0.05 compared with total SCFA levels of D-IBS, M-IBS, and HC.

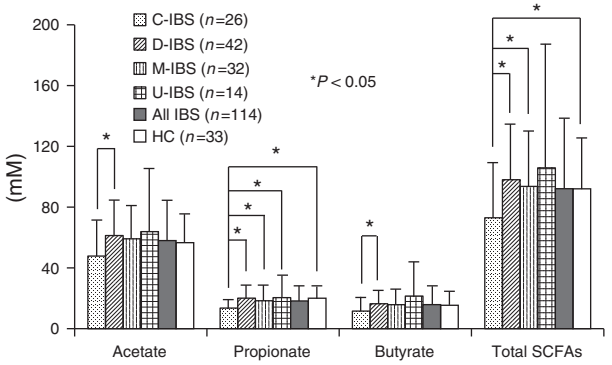


Figure 2. Fecal short-chain fatty acids (SCFAs) levels by irritable bowel syndrome (IBS) subtypes. Total SCFA level was significantly lower in constipation-predominant irritable bowel syndrome (C-IBS) than in diarrhea-predominant IBS (D-IBS), mixed type IBS (M-IBS), and healthy control (HC). Total SCFA level averaged across all IBS was similar with that in HC. Propionate level in C-IBS was significantly lower than that of the other groups. Acetate and butyrate levels were significantly lower in C-IBS than in D-IBS. **P*<0.05. U-IBS, unspecified type IBS.

Intestinal transit time

Intestinal transit time data were collected from 47 patients with IBS and 10 HCs. Overall the average intestinal transit times including SBTT, CTT, and WGTT were not significantly different between IBS patients and HC. However, as expected, CTT and WGTT were significantly longer in C-IBS compared with those in D-IBS and M-IBS. There were no significant differences in SBTT between the groups. The whole-gut and segmental transit times are presented in **Table 5**.

Correlation between colonic pH, fecal SCFAs, CTT, and symptoms

Colonic pH levels positively correlated with CTT (Spearman's correlation coefficient, SCC=0.33, *P*=0.013) and had a negative correlation trend with fecal SCFA levels (SCC=−0.23, *P*=0.079). Acetate, propionate, butyrate, and total SCFA levels correlated negatively with CTT and positively with stool frequency (total SCFAs vs. CTT, SCC=−0.44, *P*=0.001; total SCFAs vs. stool frequency, SCC=0.45, *P*<0.001).

Colonic pH levels positively correlated with IBS-SSS (SCC=0.49, *P*=0.014), and similar positive trends were noted for abdominal pain and bloating with negative trends for IBS-QOL. CTT had correlation trend positively with IBS symptom scores and negatively with IBS-QOL. Fecal SCFA levels had no correlation with IBS symptom scores and QOL. The results of Spearman's correlation analyses are presented in **Table 6**.

Table 5. Gastrointestinal transit times according to IBS subtype

	C-IBS (*n*=13)	D-IBS (*n*=14)	M-IBS (*n*=12)	U-IBS (*n*=8)	IBS total (*n*=47)	HC (*n*=10)	*P* value[a]
WGTT (hours)	57.0±27.9[b]	30.1±18.4	30.1±14.4	45.3±25.2	39.6±24.0	48.2±31.9	0.344
SBTT (hours)	4.4±2.1	4.1±1.1	3.7±1.2	3.9±1.1	4.0±1.4	3.6±1.3	0.348
CTT (hours)	44.6±24.7[c]	22.4±16.0	23.1±14.7	39.1±24.6	31.1±21.6	34.7±21.1	0.632

C-IBS, constipation-predominant IBS; CTT, colon transit time; D-IBS, diarrhea-predominant IBS; HC, healthy control; IBS, irritable bowel syndrome; M-IBS, mixed type IBS; SBTT, small bowel transit time; U-IBS, unspecified type IBS; WGTT, whole-gut transit time.
[a]*P*-value for comparison between total IBS and HC.
[b]*P*<0.05 compared with WGTTs of D-IBS and M-IBS.
[c]*P*<0.05 compared with CTTs of D-IBS and M-IBS.

Table 6. Spearman's correlation coefficient in correlation analyses between variables

	Total colon pH	Total SCFAs	CTT	Stool consistency	Stool frequency	Abdominal pain	Bloating	IBS-SSS	IBS-QOL
Total colon pH	1.00	–0.23	0.33[a]	–0.30	–0.10	0.29	0.29	0.49[a]	–0.35
Total SCFAs	–0.23	1.00	–0.44[b]	–0.01	0.45[b]	–0.14	0.01	–0.33	0.12
CTT	0.33[a]	–0.44[b]	1.00	–0.64	–0.40	0.41	0.30	0.39	–0.28

CTT, colon transit time; IBS-SSS, IBS-symptom severity scale score; IBS-QOL, IBS-quality-of-life; SCFAs, short-chain fatty acids.
[a]*P*<0.05.
[b]*P*<0.01.

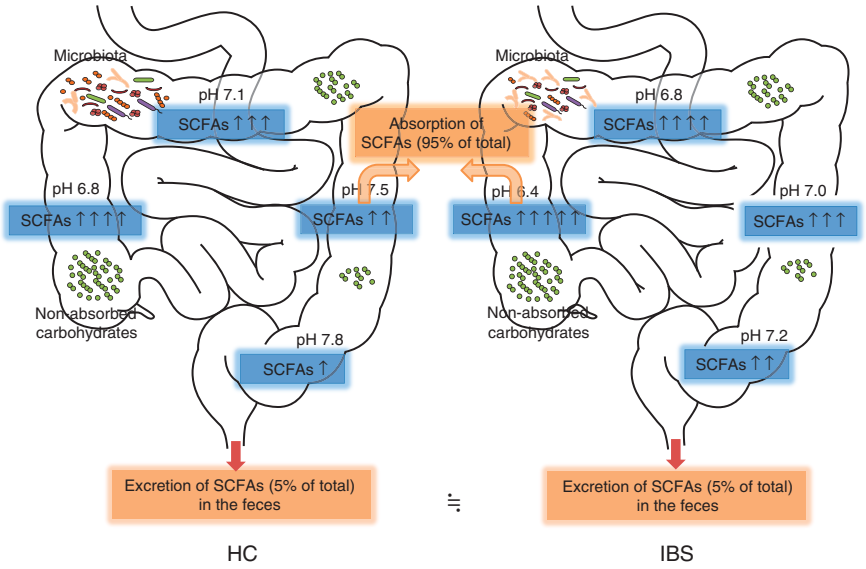


Figure 3. Hypothesized model. Colonic intraluminal pH is decreased, and short-chain fatty acid (SCFA) production is increased, suggesting higher fermentation in patients with irritable bowel syndrome (IBS) compared with healthy control (HC). SCFAs are effectively absorbed (>95% of SCFAs) in the gastrointestinal (GI) tract, and only a small portion (~5% of total SCFAs) is excreted in feces. Both processes can affect and are affected by intestinal transit.

DISCUSSION

IBS is a common condition that affects nearly 10% of the general population (1,2) and contributes to a significant decrease in QOL along with an increase in health-care utilization. In spite of its high prevalence and impact on QOL, the pathophysiological mechanisms underlying the disease remain unknown. Recently, the possible role of the intestinal microbiota in the pathogenesis of IBS has gained much interest (4,8,26). In this study, we investigated the role of intestinal bacterial fermentation in IBS.

© 2015 by the American College of Gastroenterology

Intraluminal intestinal fermentation by colonic bacteria produces gases such as hydrogen and carbon dioxide and SCFAs as secondary by-products (10,27,28). The predominant SCFAs produced by bacterial fermentation are acetate, butyrate, and propionate. Lactate, which is a precursor for propionate and butyrate, is itself a by-product of fermentation. SCFAs are naturally acidic and cause a measurable drop in the pH of the intestinal lumen (28–30). In this study, we have measured intraluminal pH levels and fecal SCFA concentrations as surrogate markers for intraluminal intestinal fermentation and used these factors to assess the location and magnitude of intestinal bacterial fermentation in a well-characterized cohort of patients with IBS and HC.

We found that the mean total colonic pH levels were significantly lower in IBS compared with HC, potentially indicating higher intraluminal bacterial fermentation in the IBS group. Of great importance is our finding of consistently higher bacterial fermentation in the large bowel of patients with IBS compared with HC, whereas there were no differences in the pH levels in the small bowel between those two groups. Notably, these findings were consistent in all small bowel and large bowel segments and across all IBS subtypes.

The intraluminal pH levels were lower at the proximal colon than the distal colon in both IBS and HC groups when we divided the quartiles based on the transit time. However, the quartiles based on transit time might not be the same as those based on geography, because if some patients might have mixtures of very rapid transit through one part of the colon and then very slow transit through the next, although these two segments might differ pathophysiologically. Although those cases were not considered common findings, it may be a limitation of our study.

Small intestinal bacterial overgrowth (SIBO) is a condition in which an excess of colonic bacteria colonize the small intestine. The symptoms of SIBO and IBS overlap (31), and it has been suggested that the prevalence of SIBO is increased in IBS or that SIBO may have an etiological role in the pathogenesis of IBS, particularly in view of the improvement of symptoms with antibiotic treatment (32–35). However, the link between SIBO and IBS is still unclear, and there are some controversies regarding the real prevalence and the importance of SIBO in the pathogenesis of IBS (36). Our study findings indicate an increased bacterial fermentation in the large but not in the small bowel of patients with IBS, thus supporting the hypothesis of altered microbial composition and/or function in the colon rather than in the small bowel (e.g., in SIBO) in the pathogenesis of IBS. In addition, our finding of significant differences in total and segmental CTT, but not in SBTT between IBS and its subtypes and HC, further supports the importance of altered colonic rather than small bowel physiology in the pathogenesis of the disorder.

The importance of SCFAs as possible etiological factors in the pathogenesis of IBS is suggested by observations from animal and human studies. Animal studies demonstrate that SCFA can initiate high-amplitude propagated colonic contractions (37) and accelerated intestinal transit and motility via intestinal release of 5-hydroxytryptamine (38). Intracolonic infusion of 0.5% acetic acid enhanced sensitivity to colorectal distension (39). In a human study, the total concentration of SCFAs in jejunal secretion was approximately four times higher in patients with SIBO than in healthy subjects (40). In addition, a recent study reported higher counts of acetic- and propionic-acid producing bacteria (*Veillonella* and *Lactobacillus*) in IBS patients and an association between the higher levels of these SCFAs and GI symptoms and QOL (41). However, the studies in this area have generated conflicting findings on fecal organic acids. For example, a study conducted in patients with D-IBS showed that fecal SCFAs were decreased compared with HC (42). In our study, there were no significant differences in the mean levels of fecal SCFAs between IBS (all subtypes) and HC, and the levels of fecal SCFAs did not correlate with IBS symptom severity. However, we were able to demonstrate that fecal SCFAs levels in C-IBS patients were significantly lower than those in D-IBS patients and tended to be lower than those in all the other IBS subtypes and HC. In addition, as expected, SCFAs levels negatively correlated with CTT, which is consistent with our finding of a positive correlation between colonic pH and CTT, indicating that higher fermentation with higher production of SCFA is associated with increased motility and shorter transit times. An alternative explanation can be reduced absorption of SCFAs in patients with rapid intestinal transit due to a shorter transit time. The difficulties in interpretation of these findings reflect the multifactorial determinants of fecal SCFAs. Intraluminal SCFAs are very efficiently absorbed in the colon, and only 5 to 10% of the SCFAs produced by bacterial fermentation are excreted and can be measured in the stool (43–45). The correlation between fecal SCFAs levels and CTT and the differences between IBS subtypes suggest that fecal SCFAs levels may be controlled by CTT more dominantly than fermentation degree; thus, fecal SCFAs may not be a sensitive marker to estimate intraluminal bacterial fermentation.

In addition to magnitude and location, we examined the associations between intestinal fermentation and clinical presentation of IBS. Unlike what we expected, the colonic pH levels positively correlated with IBS symptom scores including IBS-SSS. This finding suggests that other factors such as abnormal intestinal motility and psychological disturbances may be more important than bacterial fermentation in determining symptoms severity.

The wireless motility SmartPill capsule measures pH, pressure, and temperature in real-time throughout the GI tract (21). Overall, pH profiles in the GI tract are characterized by an abrupt rise in pH from acid to near neutral as the capsule exited the stomach, a slow continued rise in pH through the small bowel until reaching the large bowel where pH falls more than 1 unit, and subsequently there is a slow rise in pH through the colon (21,46). The pH drop at the ileocecal junction is well documented in studies using ingestible radiotelemetry capsules (47,48). Time between an abrupt rise of pH in duodenum and a fall of >1 pH unit in ileocolonic junction was taken as SBTT. CTT was assessed by measuring the time required from cecal entry to body exit (loss of signal and/or an abrupt temperature drop) (21). A recent scintigraphic study using a radioisotope-labeled WMC has validated this pH change at the ileocolonic junction and has shown that the fall in pH observed with WMC corresponds to the time of arrival of the WMC into the cecum or ascending colon (49). Several studies have confirmed the utility of WMC in quantifying CTT when compared with

the radiopaque marker method. There was a good agreement between the WMC and radiopaque marker transit results (22,46). The WMC is a validated test; however, there are some confounding issues with the test. The pH drop at the ileocecal junction is occasionally (<5%) not clearly identifiable, and device failure is reported in <3% of cases. In our study, the pH landmarks were well identified in all cases, and there were no capsule failures.

To our knowledge, there has been no previous study to evaluate bacterial fermentation in the whole intestine by assessing intraluminal pH changes in patients with IBS. However, our study may have some limitations. First, although subjects were asked to keep their regular diet during the study, we did not have a detailed dietary data and could not control for dietary effects on intestinal physiology and fermentation processes including intraluminal pH and SCFA production. Second, we did not investigate the intestinal bacterial community, so we do not know which bacterial species or genera were altered and how these alterations relate to the variables of interest. Third, we did not evaluate gas production, such as hydrogen and methane, which might be associated with symptom generation. Finally, the analyses of the wireless motility capsule (SmartPill) data included a relatively small number of subjects and involved some multiple comparisons, particularly in the sub-analysis by IBS subtypes. Further large-scale studies, including the analyses of dietary factors, gas production, and bacterial communities are needed. These studies may enable to associate alterations in bacterial fermentation with specific diet or bacterial communities and provide insights on how these associations relate to the altered physiology and symptoms observed in IBS.

In conclusion, we found that colonic intraluminal pH is decreased, suggesting increased bacterial fermentation, in IBS patients across all subtypes compared with HC, and that this process is more prominent in the proximal colon. The altered intestinal fermentation and transit in IBS appear to be in the colon and not in the small bowel. Higher levels of fecal SCFAs and lower levels of intraluminal pH are associated with decreased CTT. The hypothesized model of our study results was illustrated in the **Figure 3**. Our study further supports the notion that altered colonic bacterial fermentation has an important etiological role for in the pathogenesis of IBS. Future studies should investigate the degree to which altered intestinal fermentation is a cause of IBS symptoms or an effect of IBS-related altered physiology by, if possible, direct measurement of bacterial fermentation and clarify the possible clinical benefit of targeting intestinal fermentation in the treatment of functional GI disorders.

ACKNOWLEDGMENTS

We thank Ms Jenny C. Martin for her contribution to the SCFA analyses.

CONFLICT OF INTEREST

Guarantor of the article: Yehuda Ringel, MD.

Specific author contributions: Tamar Ringel-Kulka and Yehuda Ringel planned and performed the study, interpreted the results, and reviewed and edited the manuscript; Daniel Temas collected and arranged the data; Chang Hwan Choi and Ari Kim analyzed and interpreted data and drafted the manuscript; Joseph A Galanko provided statistical support; Karen Scott contributed to the analysis of fecal SCFAs; Daniele M Maier contributed to the wireless motility capsule (SmartPill) studies; all authors have read and approved the final draft of this paper.

Financial support: This work was supported by grants K23 DK075621 (YR) and P30 DK34987 (JAG). The project described was supported by the National Center for Advancing Translational Sciences (NCATS) and National Institutes of Health, through Grant Award Number 1UL1TR001111. KS receives support from the Scottish Government (RESAS).

Potential competing interests: None.

Study Highlights

WHAT IS CURRENT KNOWLEDGE

- ✓ Alterations in the gut microbiota have a role in irritable bowel syndrome (IBS).
- ✓ Alterations in intestinal fermentation lead to certain physiological abnormalities such as intraluminal excessive gas production and altered motility.
- ✓ Diet low in fermentation substrates such as poorly absorbable, highly fermentable carbohydrates offers considerable symptom relief in IBS.
- ✓ The understanding of the role of altered intestinal fermentation, including its magnitude and location, in the pathogenesis of IBS is not completely understood.

WHAT IS NEW HERE

- ✓ Colonic intraluminal pH decreased, indicating higher intestinal fermentation, in IBS patients across all subtypes compared with healthy controls.
- ✓ The intestinal fermentation in IBS is altered in the colon but not in the small bowel.
- ✓ Higher levels of fecal SCFAs and lower levels of intraluminal pH are associated with decreased colon transit time.

REFERENCES

1. Drossman DA, Camilleri M, Mayer EA *et al.* AGA technical review on irritable bowel syndrome. Gastroenterology 2002;123:2108–31.
2. Spiller R, Aziz Q, Creed F *et al.* Guidelines on the irritable bowel syndrome: mechanisms and practical management. Gut 2007;56:1770–98.
3. Barbara G, Cremon C, Carini G *et al.* The immune system in irritable bowel syndrome. J Neurogastroenterol Motil 2011;17:349–59.
4. Camilleri M. Peripheral mechanisms in irritable bowel syndrome. N Engl J Med 2012;367:1626–35.
5. Camilleri M, Lasch K, Zhou W. Irritable bowel syndrome: methods, mechanisms, and pathophysiology. The confluence of increased permeability, inflammation, and pain in irritable bowel syndrome. Am J Physiol Gastrointest Liver Physiol 2012;303:G775–85.
6. Carroll IM, Ringel-Kulka T, Keku TO *et al.* Molecular analysis of the luminal- and mucosal-associated intestinal microbiota in diarrhea-predominant irritable bowel syndrome. Am J Physiol Gastrointest Liver Physiol 2011;301:G799–G807.
7. Carroll IM, Ringel-Kulka T, Siddle JP *et al.* Alterations in composition and diversity of the intestinal microbiota in patients with diarrhea-predominant irritable bowel syndrome. Neurogastroenterol Motil 2012;24:521–30.e248
8. Lee BJ, Bak YT. Irritable bowel syndrome, gut microbiota and probiotics. J Neurogastroenterol Motil 2011;17:252–66.
9. Ringel Y, Maharshak N. Intestinal microbiota and immune function in the pathogenesis of irritable bowel syndrome. Am J Physiol Gastrointest Liver Physiol 2013;305:G529–41.

© 2015 by the American College of Gastroenterology

10. Chassard C, Dapoigny M, Scott KP *et al.* Functional dysbiosis within the gut microbiota of patients with constipated-irritable bowel syndrome. Aliment Pharmacol Ther 2012;35:828–38.
11. Dapoigny M, Stockbrugger RW, Azpiroz F *et al.* Role of alimentation in irritable bowel syndrome. Digestion 2003;67:225–33.
12. King TS, Elia M, Hunter JO. Abnormal colonic fermentation in irritable bowel syndrome. Lancet 1998;352:1187–9.
13. Brenner DM, Moeller MJ, Chey WD *et al.* The utility of probiotics in the treatment of irritable bowel syndrome: a systematic review. Am J Gastroenterol 2009;104:1033–49.
14. Pimentel M, Lembo A, Chey WD *et al.* Rifaximin therapy for patients with irritable bowel syndrome without constipation. N Engl J Med 2011;364:22–32.
15. Halmos EP, Power VA, Shepherd SJ *et al.* A diet low in FODMAPs reduces symptoms of irritable bowel syndrome. Gastroenterology 2014;146: 67–75. e5.
16. Staudacher HM, Lomer MC, Anderson JL *et al.* Fermentable carbohydrate restriction reduces luminal bifidobacteria and gastrointestinal symptoms in patients with irritable bowel syndrome. J Nutr 2012;142:1510–8.
17. Veldhuyzen van Zanten SJ, Talley NJ, Bytzer P *et al.* Design of treatment trials for functional gastrointestinal disorders. Gut 1999;45 Suppl 2:II69–II77.
18. Francis CY, Morris J, Whorwell PJ. The irritable bowel severity scoring system: a simple method of monitoring irritable bowel syndrome and its progress. Aliment Pharmacol Ther 1997;11:395–402.
19. Drossman DA, Patrick DL, Whitehead WE *et al.* Further validation of the IBS-QOL: a disease-specific quality-of-life questionnaire. Am J Gastroenterol 2000;95:999–1007.
20. Lewis SJ, Heaton KW. Stool form scale as a useful guide to intestinal transit time. Scand J Gastroenterol 1997;32:920–4.
21. Rao SS, Camilleri M, Hasler WL *et al.* Evaluation of gastrointestinal transit in clinical practice: position paper of the American and European Neurogastroenterology and Motility Societies. Neurogastroenterol Motil 2011;23:8–23.
22. Rao SS, Kuo B, McCallum RW *et al.* Investigation of colonic and whole-gut transit with wireless motility capsule and radiopaque markers in constipation. Clin Gastroenterol Hepatol 2009;7:537–44.
23. Chassard C, Scott KP, Marquet P *et al.* Assessment of metabolic diversity within the intestinal microbiota from healthy humans using combined molecular and cultural approaches. FEMS Microbiol Ecol 2008;66:496–504
24. Matulova M, Nouaille R, Capek P *et al.* Degradation of wheat straw by *Fibrobacter succinogenes* S85: a liquid- and solid-state nuclear magnetic resonance study. Appl Environ Microbiol 2005;71:1247–53.
25. Richardson A, Calder A, Srewart C *et al.* Simultaneous determination of volatile and non-volatile acidic fermentation products of anaerobes by capillary gas chromatography. Lett Appl Microbiol 1989;9:5–8.
26. Simren M, Barbara G, Flint HJ *et al.* Intestinal microbiota in functional bowel disorders: a Rome foundation report. Gut 2013;62:159–76.
27. Hamer HM, De Preter V, Windey K *et al.* Functional analysis of colonic bacterial metabolism: relevant to health? Am J Physiol Gastrointest Liver Physiol 2012;302:G1–G9.
28. Wong JM, de Souza R, Kendall CW *et al.* Colonic health: fermentation and short chain fatty acids. J Clin Gastroenterol 2006;40:235–43.
29. Eswaran S, Muir J, Chey WD. Fiber and functional gastrointestinal disorders. Am J Gastroenterol 2013;108:718–27.
30. Layden BT, Angueira AR, Brodsky M *et al.* Short chain fatty acids and their receptors: new metabolic targets. Transl Res 2013;161:131–40.
31. Pimentel M, Soffer EE, Chow EJ *et al.* Lower frequency of MMC is found in IBS subjects with abnormal lactulose breath test, suggesting bacterial overgrowth. Dig Dis Sci 2002;47:2639–43.
32. Majewski M, McCallum RW. Results of small intestinal bacterial overgrowth testing in irritable bowel syndrome patients: clinical profiles and effects of antibiotic trial. Adv Med Sci 2007;52:139–42.
33. Majewski M, Reddymasu SC, Sostarich S *et al.* Efficacy of rifaximin, a nonabsorbed oral antibiotic, in the treatment of small intestinal bacterial overgrowth. Am J Med Sci 2007;333:266–70.
34. Pimentel M, Chow EJ, Lin HC. Eradication of small intestinal bacterial overgrowth reduces symptoms of irritable bowel syndrome. Am J Gastroenterol 2000;95:3503–6.
35. Pimentel M, Chow EJ, Lin HC. Normalization of lactulose breath testing correlates with symptom improvement in irritable bowel syndrome. a double-blind, randomized, placebo-controlled study. Am J Gastroenterol 2003;98:412–9.
36. Yu D, Cheeseman F, Vanner S. Combined oro-caecal scintigraphy and lactulose hydrogen breath testing demonstrate that breath testing detects oro-caecal transit, not small intestinal bacterial overgrowth in patients with IBS. Gut 2011;60:334–40.
37. Kamath PS, Hoepfner MT, Phillips SF. Short-chain fatty acids stimulate motility of the canine ileum. Am J Physiol 1987;253:G427–33.
38. Fukumoto S, Tatewaki M, Yamada T *et al.* Short-chain fatty acids stimulate colonic transit via intraluminal 5-HT release in rats. Am J Physiol Regul Integr Comp Physiol 2003;284:R1269–76.
39. Winston J, Shenoy M, Medley D *et al.* The vanilloid receptor initiates and maintains colonic hypersensitivity induced by neonatal colon irritation in rats. Gastroenterology 2007;132:615–27.
40. Hoverstad T, Bjorneklett A, Fausa O *et al.* Short-chain fatty acids in the small-bowel bacterial overgrowth syndrome. Scand J Gastroenterol 1985;20:492–9.
41. Tana C, Umesaki Y, Imaoka A *et al.* Altered profiles of intestinal microbiota and organic acids may be the origin of symptoms in irritable bowel syndrome. Neurogastroenterol Motil 2010;22:512–9. e114-5.
42. Treem WR, Ahsan N, Kastoff G *et al.* Fecal short-chain fatty acids in patients with diarrhea-predominant irritable bowel syndrome: in vitro studies of carbohydrate fermentation. J Pediatr Gastroenterol Nutr 1996;23:280–6.
43. Roediger WE, Moore A. Effect of short-chaim fatty acid on sodium absorption in isolated human colon perfused through the vascular bed. Dig Dis Sci 1981;26:100–6.
44. Ruppin H, Bar-Meir S, Soergel KH *et al.* Absorption of short-chain fatty acids by the colon. Gastroenterology 1980;78:1500–7.
45. McNeil NI, Cummings JH, James WP. Short chain fatty acid absorption by the human large intestine. Gut 1978;19:819–22.
46. Camilleri M, Thorne NK, Ringel Y *et al.* Wireless pH-motility capsule for colonic transit: prospective comparison with radiopaque markers in chronic constipation. Neurogastroenterol Motil 2010;22:874–82. e233.
47. Evans DF, Pye G, Bramley R *et al.* Measurement of gastrointestinal pH profiles in normal ambulant human subjects. Gut 1988;29:1035–41.
48. Watson BW, Meldrum SJ, Riddle HC *et al.* pH profile of gut as measured by radiotelemetry capsule. Br Med J 1972;2:104–6.
49. Zarate N, Mohammed SD, O'Shaughnessy E *et al.* Accurate localization of a fall in pH within the ileocecal region: validation using a dual-scintigraphic technique. Am J Physiol Gastrointest Liver Physiol 2010;299:G1276–86.

NEUROGASTROENTEROLOGY

Neuronal correlates of placebo in chronic FGIDs

QiQi Zhou and G. Nicholas Verne

Placebo analgesia is increasingly appreciated in many difficult to treat chronic functional gastrointestinal disorders such as IBS. However, investigations of interactions between psychological and biological placebo factors are still in early stages. Now, technologies have been developed that enable neural mechanisms of placebo analgesia to be studied more directly in humans.

Zhou, Q. & Verne, G. N. *Nat. Rev. Gastroenterol. Hepatol.* **11**, 460–461 (2014); published online 8 July 2014; doi:10.1038/nrgastro.2014.118

IBS is one of the most common functional gastrointestinal disorders (FGIDs) in which the underlying pathophysiology of the disorder remains poorly understood. Some patients with IBS have heighted perception to both somatic and visceral stimuli, suggesting abnormalities in central and/or peripheral pain processing (a theory supported by neuroimaging studies that reveal altered brain activation in response to nociceptive stimuli).[1–4] This ongoing sensitization leads to altered input to the central neuroaxis, which might induce long-term microstructural reorganization of the brain and subsequent failure of conventional therapies and a robust placebo response in patients with IBS.[5] A new study by Schmid *et al.*[6] sought to further understand the neural circuitry of placebo analgesia.

Schmid and colleagues[6] compared nociceptive rectal distension in patients with IBS, patients with ulcerative colitis and healthy individuals as controls after administration of a saline placebo with either instructions of pain relief (placebo; patients were told they were receiving a highly potent analgesic drug when in fact they were administered saline) or neutral instructions (control; patients were truthfully told that saline was administered). Neural activation as assessed by functional MRI (fMRI), which was evaluated along with ratings of perceived and expected pain, as well as measures of salivary cortisol concentrations and negative affectivity. Patients with IBS did not exhibit neural down-regulation of rectal-distension-induced pain during placebo analgesia, whereas controls and patients with ulcerative colitis both demonstrated marked reductions in neuronal activity in the cingulate and somatosensory cortex. Interestingly, patients with IBS displayed an enhanced activation of the somatosensory cortex in the placebo condition. Psychological testing further revealed that depression scores correlated with a weaker placebo analgesia.

Much progress has been made over the past century in characterizing the neural circuitry that controls inhibition of pain.[3] Ascending and descending pain pathways terminate within cortical and subcortical structures that are the basis for pain-modulatory pathways such as the amygdala, thalamus, hypothalamic nuclei, rostroventral medulla and midbrain periaqueductal grey (Figure 1). These structures are central targets of transmission and modulation of nociceptive pain pathways. Intact descending inhibitory modulation of nociceptive afferent impulses is paramount to enable adequate pain control in many chronic pain disorders such as IBS. Several recent technologies and experimental approaches have been developed whereby neural mechanisms of analgesia can be studied more directly in human study participants. These approaches include functional imaging methods such as PET or fMRI. One or several of these measurements can be combined with valid measures of expectancy and desire for pain relief, as well as with reliable measurement of pain intensity and pain unpleasantness. Such an approach makes it possible to

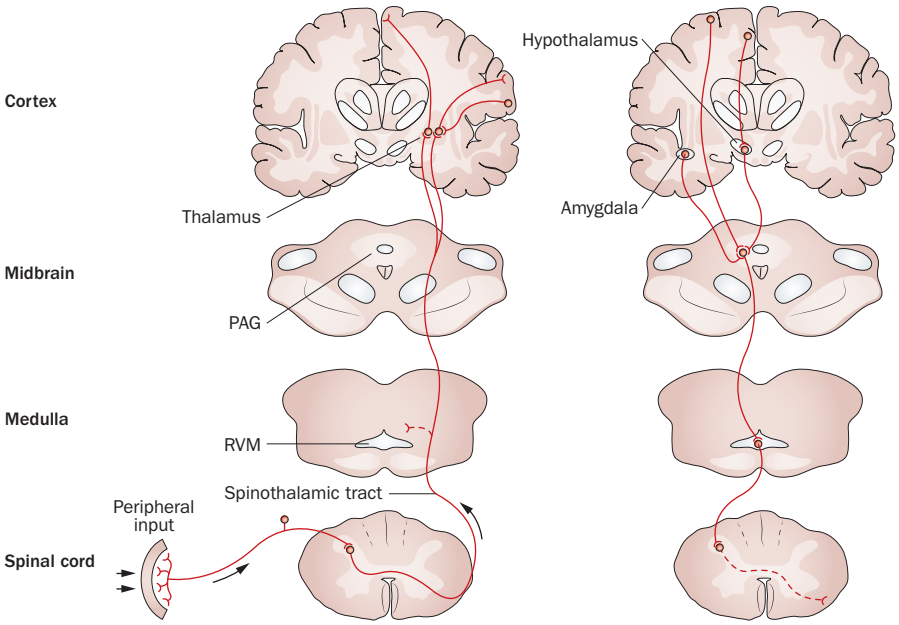


Figure 1 | Pain-modulatory pathways. This diagram depicts ascending (left) and descending (right) pain pathways (red lines) that terminate within cortical and subcortical structures and are the basis for pain-modulatory pathways and include the amygdala, hypothalamic nuclei or hypothalamus, thalamus, midbrain PAG, and the RVM. These cortical and subcortical structures are central targets of transmission and modulation of nociceptive pain pathways. Abbreviations: PAG, periaqueductal grey; RVM, rostroventral medulla.

© 2014 Macmillan Publishers Limited. All rights reserved

simultaneously examine how the placebo analgesic effect is expressed at different levels. The potential for analysis of relationships between cognitive mediating factors and cerebral pain-modulating circuitry in the case of placebo analgesia was first illuminated by two early studies.

The first study used neuroimaging of displacement of a radioactive opioid ligand ([^{11}C] carfentanil) to examine the function of the opioid system and μ-opioid receptors of the brains of healthy humans undergoing sustained pain.[7] Sustained pain induced regional release of endogenous opioids in brain regions known to modulate pain: the rostral anterior cingulate cortex; ipsilateral amygdala; and contralateral thalamus. Activation of the μ-opioid receptor system at all of these sites was associated with reductions in the sensory and affective ratings of pain. A second study used PET to demonstrate that both opioid and placebo analgesia are associated with increased activity in the rostral anterior cingulate cortex in healthy humans.[8] They also observed a covariation between activity in the rostral anterior cingulate cortex and the lower brainstem during both exogenous opioid and placebo analgesia, but not during pain alone. Both investigations suggest the feasibility of relating psychological mediating factors of placebo analgesia to brain regions involved in pain modulation.

> “Some patients with IBS have heighted perception to both somatic and visceral stimuli...”

The possibility of a refined analysis of placebo effects within studies of patients with FGIDs has far-reaching scientific and medical implications.[9] Our present limited capacity to ascertain, measure and control for placebo effects is at the heart of complex and difficult questions about pharmacological therapies for pain as well as many nonpharmacological therapies, particularly those related to surgery, hypnosis, electrical stimulation and alternative complementary and medical treatments. One potential improvement for investigations into the placebo effect would be to include a clinically relevant type of stimulation that would be more likely to activate at least some of the mechanisms believed to be important in placebo analgesia than would typical experimental pain stimuli. We need to know more about how mediating factors determine the presence and magnitude of placebo analgesia, as well as about their psychological and neural mechanisms. Advances in the measurement of placebo analgesia will not only contribute to a more refined understanding of the mechanisms in placebo analgesia, but is also likely to contribute to a more precise determination of the magnitude of placebo analgesia. Indeed, the inability of patients with IBS to effectively engage neural downregulation of visceral pain signals might help to illuminate the high rate of failure of many pharmacological therapies for functional bowel pain.[6,9] Strategies can be developed whereby it is possible to assess most or all of the relevant mediating factors that contribute to placebo effects.

The latest neuroimaging studies raise several important considerations regarding the understanding of the neural mechanisms of the placebo effect in patients with FGIDs.[6,10] Interestingly, patients with IBS in the study by Schmid *et al.*[6] did not exhibit visceral hyperalgesia. A subset of patients with IBS (30–40%) exhibit visceral hypersensitivity to colonic distension,[4] and it is this group of patients that would be expected not to adequately engage neural downregulation of visceral pain. Future studies are needed to compare neural placebo responses between patients with IBS with and without visceral hypersensitivity. In addition, other pain transmission pathways need to be assessed during placebo conditioning in patients with IBS, including the hypothalamic nuclei, rostroventral medulla and the midbrain periaqueductal grey. Also, a more detailed analysis of the patients' desire and expectancy for pain relief during placebo analgesia might provide additional important information along with imaging in patients with IBS before and after cognitive–behavioural therapy that might augment descending inhibitory pain pathways. Finally, future imaging studies are needed to evaluate neuronal deactivation during placebo analgesia, which could delineate additional cerebral patterns important in the placebo response. Thus, it is possible that, in the future, placebo analgesia could be used to augment established medical therapies for FGIDs.

*Division of Gastroenterology & Hepatology, University of Texas Medical Branch, 301 University Boulevard, 4.106 McCullough Building, Rt. 0764, Galveston, TX 77555-0764, USA (**Q.Z.**, **G.N.V.**).*
Correspondence to: G.N.V.
gnverne@utmb.edu

Acknowledgements
The authors are supported by a grant from the NIH from National Institute of Diabetes and Digestive and Kidney diseases (DK099052).

Competing interests
The authors declare no competing interests.

1. Zhou, Q., Fillingim, R. B., Riley, J. L., Malarkey, W. B. & Verne, G. N. Central and peripheral hypersensitivity in the irritable bowel syndrome. *Pain* **148**, 454–461 (2010).
2. Mayer, E. A. & Tillisch, K. The brain–gut axis in abdominal pain syndromes. *Annu. Rev. Med.* **62**, 381–396 (2011).
3. Price, D. D. *et al.* Widespread hyperalgesia in irritable bowel syndrome is dynamically maintained by tonic visceral impulse input and placebo/nocebo factors: Evidence from human psychophysics, animal models, and neuroimaging. *Neuroimage* **47**, 995–1001 (2009).
4. Zhou, Q. & Verne, G. N. New insights into visceral hypersensitivity-clinical implications in IBS. *Nat. Rev. Gastroenterol. Hepatol.* **8**, 349–355 (2011).
5. Ellingson, B. M. *et al.* Diffusion tensor imaging detects microstructural reorganization in the brain associated with chronic irritable bowel syndrome. *Pain* **154**, 1528–1541(2013).
6. Schmid, J. *et al.* Placebo analgesia in patients with functional and organic abdominal pain: a fMRI study in IBS, UC and healthy volunteers. *Gut* http://dx.doi.org/10.1136/gutjnl-2013-306648.
7. Zubieta, J. K. *et al.* Regional μ-opioid receptor regulation of sensory and affective dimensions of pain. *Science* **293**, 311–315 (2001).
8. Petrovic, P., Kalso, E., Petersson, K. M. & Invar, M. Placebo and opioid analgesia —imaging a shared neuronal network. *Science* **295**, 1737–1740 (2002).
9. Hróbjartsson, A. & Gøtzsche, P. C. Is the placebo powerless? An analysis of clinical trials comparing placebo with no treatment. *N. Engl. J. Med.* **24**, 1594–1602 (2001).
10. Lee, H. F. *et al.* Enhanced affect/cognition-related brain responses during visceral placebo analgesia in irritable bowel syndrome patients. *Pain* **153**, 1301–1310 (2012).

© 2014 Macmillan Publishers Limited. All rights reserved

nature COMMUNICATIONS

ARTICLE

Received 23 Jan 2014 | Accepted 9 Jun 2014 | Published 8 Jul 2014

DOI: 10.1038/ncomms5344 OPEN

Tipping elements in the human intestinal ecosystem

Leo Lahti[1,2], Jarkko Salojärvi[1,*], Anne Salonen[3,*], Marten Scheffer[4] & Willem M. de Vos[1,2,3]

The microbial communities living in the human intestine can have profound impact on our well-being and health. However, we have limited understanding of the mechanisms that control this complex ecosystem. Here, based on a deep phylogenetic analysis of the intestinal microbiota in a thousand western adults, we identify groups of bacteria that exhibit robust bistable abundance distributions. These bacteria are either abundant or nearly absent in most individuals, and exhibit decreased temporal stability at the intermediate abundance range. The abundances of these bimodally distributed bacteria vary independently, and their abundance distributions are not affected by short-term dietary interventions. However, their contrasting alternative states are associated with host factors such as ageing and overweight. We propose that the bistable groups reflect tipping elements of the intestinal microbiota, whose critical transitions may have profound health implications and diagnostic potential.

[1] Department of Veterinary Biosciences, University of Helsinki, PO Box 66, FI-00014 Helsinki, Finland. [2] Laboratory of Microbiology, Wageningen University, PO Box 8033, 6700 EJ Wageningen, The Netherlands. [3] Department of Bacteriology and Immunology, Immunobiology Research Program, Haartman Institute, University of Helsinki, PO Box 21, FI-00014 Helsinki, Finland. [4] Aquatic Ecology, Wageningen University, PO Box 47, 6700 AA Wageningen, The Netherlands. * These authors contributed equally to this work. Correspondence and requests for materials should be addressed to L.L. (email: leo.lahti@iki.fi).

© 2014 Macmillan Publishers Limited. All rights reserved.

The versatile and dynamic microbial ecosystems inhabiting our intestinal tract can have profound impact on well-being and health[1–4]. Although the compositional and functional properties of the intestinal microbiota have been studied extensively, we have limited understanding of the mechanisms that control this complex ecosystem[5,6]. A fundamental question is whether the overwhelming variability of the microbiota composition occurs along smooth phylogenetic gradients or through abrupt switches between alternative stable states[7]. In ecology, the concept of alternative stable states separated by tipping points has emerged as a theoretical framework with major practical implications[8].

Here we demonstrate the same properties in the human intestinal ecosystem. While previous studies have focused on analysing ecosystem-wide configurations in human intestinal microbiota[7,9,10], our analysis of a thousand western adults shows that there are distinct groups of bimodally distributed bacteria that tend to be either very abundant or nearly absent, and exhibit reduced stability at the intermediate abundance range between these two contrasting states. We propose that the bistable groups we identified represent 'tipping elements'[11], specific components of the intestinal microbiota that exhibit alternative stable states linked to the overall ecosystem state and our physiology. While further studies are needed to establish the possible causal relations, the bistable taxa can serve as indicators of the community state and its potential health associations, and resetting these 'bacterial switches' may be a radically new way to approach the rapidly growing number of health issues related to the intestinal microbiota, changing the way we look at management of the intestinal ecosystem.

Results

Identification of alternative states. To systematically investigate the hypothesis of alternative stable states in the intestinal microbiota, we analysed 130 genus-like phylogenetic groups that cover the majority of the known bacterial diversity of the human intestine[12] (Supplementary Table 1). Our analysis is based on >2,000 standardized phylogenetic microarray hybridizations that represent the faecal microbiota of a thousand adults ($n = 1{,}006$; 18–77 years) from 15 western countries (Europe and the United States; see Methods). On the basis of this extensive data collection, we identified several bacterial groups exhibiting robust bimodal abundance distributions (Fig. 1; Supplementary Table 2), including *Dialister* spp., relatives of *Bacteroides fragilis*, *Prevotella melaninogenica*, *P. oralis* and two groups of the uncultured Clostridiales (UCI and UCII; Supplementary Note 1). The strongly correlated *P. melaninogenica* and *P. oralis* were combined into a single *Prevotella* group (Supplementary Note 2). The other 124 genus-like groups that we investigated were either rare or exhibited symmetric or skewed abundance types (Fig. 2; Supplementary Table 1).

To detect alternative states in the intestinal bacteria, we used potential analysis that was recently applied to study analogous state shifts in climatology and forest ecosystems[13,14]. This theoretical framework for state identification in stochastic dynamical systems is well suited for studying the remarkably variable intestinal microbiota. We used bootstrap analysis to ensure that the findings were robust to random sampling variations, and selected for further analysis the taxonomic groups that showed consistent evidence for multimodality with a bootstrap support of ≥98% for *Dialister* spp., *P. oralis*, *P. melaninogenica*, UCI and UCII groups, and 93% for the *B. fragilis* group. The evidence was the strongest for two alternative states in all these groups, and further supported by an alternative approach, the prediction strength index (PSI)[10,15], which provided strong[10] evidence of bimodality (PSI > 0.90) for *Dialister* spp., UCI, UCII and relatives of *P. melaninogenica* and *P. oralis*, and a moderate support for the *B. fragilis* group (PSI > 0.80).

To control the possibility that the bimodal distributions could be explained by diet or other independent environmental factors due to potentially unbalanced sampling, we confirmed the bimodality in distinct subsets of the data collection, controlled potential cross-hybridization effects between closely related taxa (Supplementary Note 2) and investigated the available host and environmental parameters (see Methods). In all cases, except the *B. fragilis* group, the bimodality was supported in independent sample sets analysed with different DNA extraction methods, including two sets of samples extracted with different combination of mechanical and chemical lysis steps ($N = 401$ and $N = 287$). Moreover, the 287 samples were all from central Europe, and thus expected to be less heterogeneous regarding the long-term dietary patterns compared with the entire data collection that included subjects from the United States and from 14 countries in Europe. Since long-term habitual diets are correlated with geographic regions, the verification of bimodality in two different data sets with different geographical distributions supports our conclusions by reducing the possibility that the observed bimodality is driven by country-specific dietary or other life-style differences within the study cohort. Although all DNA preparations were identically processed for the microarray analysis, the verification in independent subsets of the data collection further helps to exclude the possibility that the detected bimodality derives from technical variation in sample processing.

Analysis of an additional data set of 138 individuals from five dietary intervention trials indicated that the observed bimodality is robust against short-term dietary interventions as we did not observe significant population-level differences in the mean abundances between the baseline and post-intervention samples in the bimodal taxa in these studies ($P > 0.05$; *t*-test; Benjamini–Hochberg correction; see Methods). Remarkably, the bimodal patterns were also indicated among the 46 placebo-treated control subjects in these controlled dietary intervention trials (Supplementary Fig. 1), further supporting the conclusion that the population-level bimodal patterns and the observed state shifts cannot be explained by short-term dietary differences. Whereas the observed states were robust to short-term dietary interventions, this does not exclude the possibility that targeted dietary interventions could be potentially used to manipulate specific bistable groups. On the basis of the available data collected in intervention trials, a pronounced dietary effect on the bimodal taxa was observed, for instance, on a formula-based weight-loss diet during which 29% (9/31) of the subjects exhibited a switch from the high to the low-abundance state of *Dialister* spp. These profound individual changes were, however, masked by remarkable within-state variation among the other subjects so that the average group-level fold change following the intervention was moderate and only borderline significant ($\log_{10}\mathrm{FC} = -0.39$; $P = 0.08$; paired Wilcoxon test). Although the observed states appear predominantly robust to dietary variation, we and others have shown that the abundance of the implicated bacteria can be to some extent affected by specific interventions as the quantity and quality of consumed carbohydrates can alter the abundance of *Bacteroides* and *Prevotella* species[16–18] as well as that of *Dialister* spp.[19] However, the reported changes were relatively small (average $|\log_{10}\mathrm{FC}| < 0.5$) in the bimodal groups that we now report[16,17] and thus induced mainly within-state variation according to the present analysis. Hence, further studies are needed to assess the possibility to induce state switches in bistable taxa based on dietary or other interventions.

© 2014 Macmillan Publishers Limited. All rights reserved.

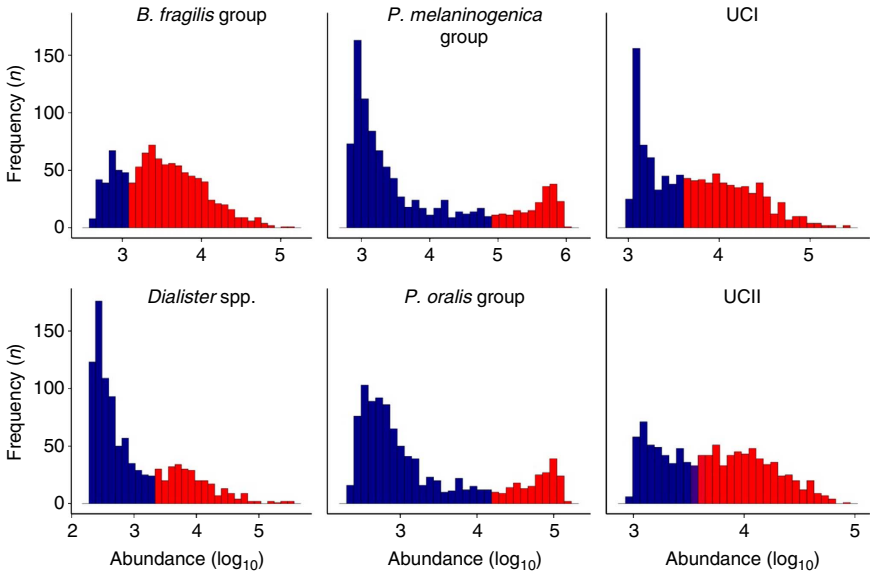


Figure 1 | The bimodal bacteria. Logarithmic abundance distributions of the six bimodal phylogenetic groups that exhibit robust alternative states of low (blue) and high (red) abundance across intestinal microbiota of 1,006 western adults. The UCI and UCII refer to the uncultured Clostridiales I and II, respectively. The frequency of the observations is shown as a function of the phylogenetic microarray $\log_{10}$ signal[12].

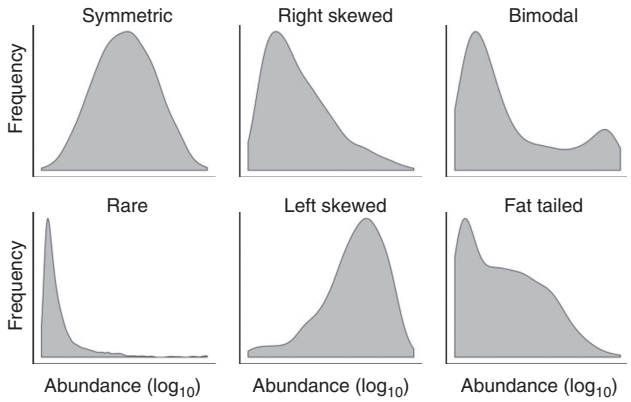


Figure 2 | Bacterial abundance types. Bacterial abundance types include symmetric, skewed and bimodal logarithmic abundance distributions (Supplementary Table 1). The skewed types include prevalent left- and right-skewed types as well as rare bacteria. The bimodal types (Fig. 1) include cases with two distinct peaks, and cases with a peak at low abundance combined with a more widely varying fat tail of high-abundance subjects. The population frequencies of the log abundance across the 1,006 western adults are shown for representative examples from each category: *Anaerostipes caccae et rel.* (symmetric); *Serratia* spp. (rare); *Streptococcus bovis et rel.* (right skewed); *Faecalibacterium prausnitzii et rel.* (left skewed); *P. oralis et rel.* (bimodal distribution with two distinct peaks); and uncultured Clostridiales I (fat tailed, bimodal distribution with a variable high-abundance state).

Temporal dynamics. The theory of stochastic dynamical systems predicts[8] that alternative stable states are separated by an intermediate unstable region, called a tipping point, where even small fluctuations may lead to an abrupt, self-propagating shift to an alternative state. While this can give rise to bimodal abundance distributions[8], as in our reported bimodal bacteria that tend to be either rare or abundant in most subjects, the observed bimodality could also be caused by other factors such as diet or yet unknown host factors, dividing the study cohort into two distinct subpopulations. Thus, bimodality of states provides only indirect evidence for bistability. In bistable systems, observations in the intermediate region between the alternative states may correspond to subjects in a transitional phase between the states[8], for instance, due to changes in diet or physiological state. We therefore looked into the temporal dynamics of the states based on a subset of 78 subjects (27–63 years) in our main cohort with additional follow-up samples within 1–9 months from the baseline (see Methods). This complements our cross-sectional analysis that represents a static snapshot of the intestinal ecosystem in each individual. A closer look at the temporal fluctuations (Fig. 3) indicated that subjects closer to the tipping point indeed had larger temporal deviations, suggesting that the intermediate abundances appear less attractive than the contrasting alternative states. We confirmed this by quantifying the stability of bacterial abundances at the intermediate abundance range. The observed deviations from the baseline in the later time points were higher in subjects whose baseline abundances were closer to the intermediate (50%) abundance range, as indicated by significant negative correlation (Spearman's coefficient ρ) between these variables for the *Dialister* spp. (-0.54), *P. melaninogenica* (-0.40), *P. oralis* (-0.34), UCI (-0.37), UCII (-0.53) and *B. fragilis* (-0.19) groups. These correlations were significant ($P<0.01$; upper tail (AS89) test; Benjamini–Hochberg adjustment) for all bimodal groups, except *B. fragilis* ($P=0.1$). Moreover, comparison between the altogether 60 prevalent taxa showed that the bimodal groups had significantly reduced intermediate stability compared with the other groups (Fig. 4; $P<0.002$; hyper-geometric test). This observation supports the hypothesis of two alternative stable states of low and high abundance divided by an unstable tipping point at the intermediate abundance range in these taxa, in line with the theoretical predictions outlined above for bistable systems[8]. Furthermore, the comparison between bimodality and intermediate stability in Fig. 4 suggested the presence of further bistable bacteria (Firmicutes), including *Clostridium* spp. (*sensu stricto*) and relatives of *C. colinum*, *Lactobacillus plantarum* and uncultured Mollicutes, as well as the *C. difficile* group that exhibits exceptionally low intermediate stability and a moderate support for bimodality, in line with earlier observations that a

© 2014 Macmillan Publishers Limited. All rights reserved.

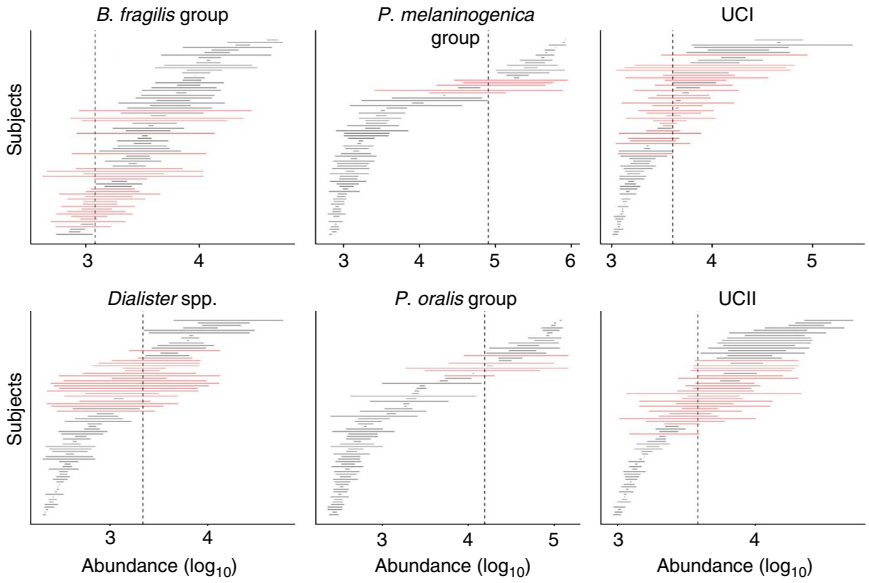


Figure 3 | Temporal variation. Temporal variation in each bimodal group during the follow-up interval (1–9 months from the baseline for 95% of the subjects, see Methods). Each horizontal line indicates the abundance range of the indicated bacteria for a single subject. The 78 subjects (horizontal lines) are ordered based on their mean abundance. The red lines highlight subjects exhibiting a state shift (Supplementary Fig. 3) across the estimated tipping point (dashed vertical lines). The differences in the follow-up times did not significantly affect the observed fluctuations ($P>0.05$; Pearson correlation; Benjamini–Hochberg adjustment).

small fraction of healthy adults carry *C. difficile*[20], one of the members of this group. Toxin-producing derivatives of *C. difficile* have been identified as the cause of severe and recurrent antibiotic-associated diarrhoea, and recent faecal transplantations have been shown to provide efficient means to manipulate its alternative states and treat the disease, resulting in complete recovery[21,22]. It is currently not known whether the alternative states of *C. difficile* as part of the normal microbiota are associated with the infection risk. While these observations suggest bistable properties for these taxa, we could not confirm their alternative states in the other tests. Furthermore, we analysed the temporal stability of the above-mentioned states, quantified as the fraction of subjects remaining in the initial state after a 3-month interval as estimated by Kaplan–Meier survival analysis (see Methods). The states with higher population frequencies were more stable in general (Supplementary Fig. 2a), in line with predictions based on the stationary-state continuous Markov process (Supplementary Fig. 3) and consistent with the idea that changes in species abundance distributions can be used to infer changes in the stability and resilience of alternative states[23]. The high-abundance *Prevotella* group was a notable exception as one of the least frequent but yet one of the most stable states. The alternative states of the reported bistable groups were also more stable than expected based on empirical distributions derived from the unimodal taxa (Supplementary Fig. 2b); the *Prevotella* group exhibited less state mixing than any other group (100%), followed by UCI (less mixing than in 93% of the comparison groups), *Dialister* spp. (83%) and UCII (78%). The *B. fragilis* group was again an exception (35%).

Analysis of individual phylotypes and methanogenic archaea. Phylotype-level analysis of the six bimodal groups indicated that the observed bistability is often associated with particular phylotypes, and absent in others (Supplementary Figs 4 and 5; Supplementary Note 3). This could indicate potential antagonism or metabolic specialization between the implicated phylotypes. In certain groups, such as the UCI and UCII, the most prevalent phylotypes exhibited stronger bimodality than their higher-level genus-like group (Supplementary Fig. 4). Of the bistable groups indicated in Fig. 4, only the *Dialister* spp., relatives of *P. melaninogenica* and *P. oralis*, and UCI/UCII groups, whose bistability was systematically supported in the different analyses, had prevalent phylotypes according to our criteria. Moreover, we observed notable variation within the genus *Prevotella* as *P. ruminicola* exhibited only moderate correlation with the combined *P. oralis*/*P. melaninogenica* group ($r=0.39$). The most strikingly bimodal bacteria were uncultured phylotypes, as in the case of the *P. melaninogenica* and *P. oralis* groups, where the type species were detected only at low abundance, but the related uncultured bacteria exhibited bimodal distributions (Supplementary Fig. 4). *D. invisus* (*Dialister* spp.) is an example of a cultured species with a strong bimodal abundance distribution (Supplementary Fig. 5).

We augmented the phylotype analysis with quantification of methanogenic archaea as this group is not targeted by the Human Intestinal Tract Chip (HITChip) phylogenetic microarray. The analysis of alternative states in methanogenic archaea is of interest as the current evidence is inconclusive regarding their prevalence in human populations. Traditionally, only a fraction of individuals have been considered as carriers, but a recent molecular study reported that the dominant archaeon, *Methanobrevibacter smithii*, is present in >95% of individuals[24]. We observed indications of bimodality in *Methanobrevibacter* spp. based on the quantitative PCR-based[25] enumeration, which was available for 53 subjects from our main cohort. On the basis of this analysis, we found 47% of these unique subjects to be carriers of *Methanobrevibacter*. Moreover, the logarithmic abundance distribution (Fig. 5) exhibited clear bimodality, representing carriers and non-carriers. Quantitative PCR-based data for these archaea were available for 452 additional samples from 189 unique subjects, of which 57% were found to be carriers. Among the 53 subjects in common with our main cohort, the abundance of these archaea had moderate positive correlation with the *Prevotella* group ($r=0.32$; $P=0.02$; t-statistics), UCI ($r=0.28$;

© 2014 Macmillan Publishers Limited. All rights reserved.

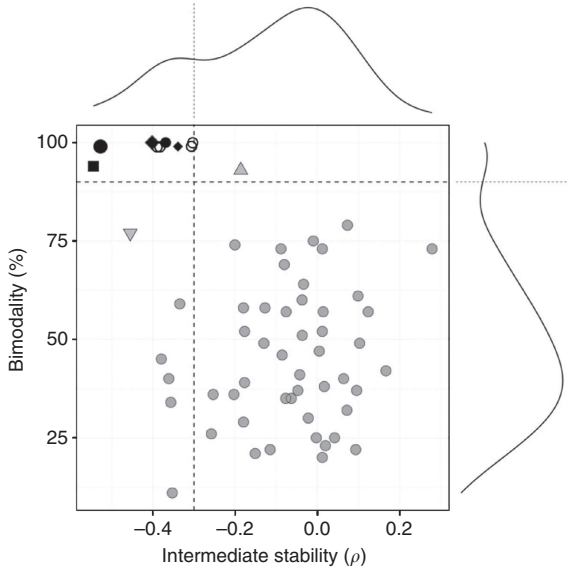


Figure 4 | Bimodality is associated with reduced intermediate stability. Analysis of the 60 prevalent taxa indicates a group of bistable bacteria with robust bimodal abundance distributions (>90% bootstrap support; vertical axis) and reduced intermediate stability ($\rho < -0.3$; horizontal axis), as also indicated by the corresponding frequency histograms (side panels). The bimodality scores are shown for the 401 samples extracted by mechanical lysis[25]; the stability estimates are based on the 78 follow-up subjects. The black symbols indicate the bistable groups whose bimodality was supported in multiple data sets, including the *Dialister* spp. (square), *P. melaninogenica* (large diamond), *P. oralis* (small diamond), UCI (small circle) and UCII (large circle) groups. These show also significantly reduced intermediate stability compared with the other prevalent groups ($\rho < -0.3$; $P < 0.002$; hyper-geometric test). The other bistable taxa (white circles) whose bimodality we could not verify in the other data sets include Firmicutes (*Clostridium* spp. (*sensu stricto*), relatives of *C. colinum*, *Lactobacillus plantarum* and uncultured Mollicutes). The bimodal *B. fragilis* group (grey upward triangle) has only moderately reduced intermediate stability ($\rho = -0.19$). The *C. difficile* group (grey downward triangle) has the third lowest intermediate stability.

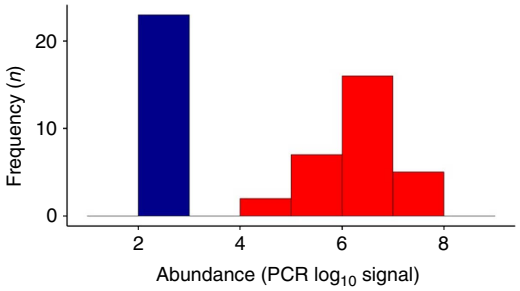


Figure 5 | Abundance distribution of *Methanobrevibacter*, the dominant methanogenic archaeon in the human intestine. Quantitative PCR-based (qPCR)[25] logarithmic abundance distribution, expressed as 16S rRNA gene copies per gram of faeces, indicates bimodality of this genus. *Methanobrevibacterium* spp. were detected in 47% of the 53 unique subjects from our main cohort for which the qPCR data were available. For subjects where *Methanobrevibacterium* spp. were not detected, the abundance is set at the detection limit of the assay. The colours indicate the two groups of subjects where the *Methanobrevibacter* spp. are absent or below the detection limit (blue), or detected (red).

$P = 0.04$) and UCII ($r = 0.23$; $P = 0.1$) groups that we also found to be associated with increased community diversity and metagenomic richness (Supplementary Fig. 6; Supplementary Note 4).

Tipping elements and the ecosystem state. While multimodality in specific microbial groups, notably the *Prevotella* genus, has previously been reported[10], our analysis complements and extends these earlier studies by providing increased phylogenetic depth and sample size, and by suggesting previously unreported taxonomic groups exhibiting bistable properties that associate to the overall intestinal ecosystem state and host health. The six bimodal genus-like groups that we identified exhibited only weak preferences for mutual co-occurrence or exclusion ($|\rho| < 0.16$ among the 401 samples with mechanical lysis; Fig. 6), except a positive correlation between UCI and UCII ($\rho = 0.50$) and between *P. melaninogenica* and *P. oralis* ($\rho = 0.97$), which we have combined into a single *Prevotella* group. The low- and high-abundance states of the resulting five bimodal groups yield altogether 32 (2^5) possible state combinations with varying population frequencies (Fig. 6), suggesting that the alternative stable states associated with distinct subcommunities often coexist in a host in different combinations. We observed, for instance, a substantial overlap between the *B. fragilis* and *Prevotella* groups, with no significant differences in *B. fragilis* abundance between the low- and high-abundance *Prevotella* subjects ($\log_{10}FC < 0.05$; *t*-test; $P > 0.05$). Similar results on the genus *Bacteroides* have been reported earlier[26], in contrast with other studies that have reported a notable trade-off between *Bacteroides* and *Prevotella* spp[6,9,27]. The bistable groups also covaried with other bacteria (Fig. 7), suggesting that their state shifts can be associated with wider changes in the overall ecosystem composition, and could be potentially used as indicators of the overall community state. Establishing possible causal relations between the bistable groups and ecosystem composition will require further studies, however.

Ecosystem-level investigation based on principal component analysis (PCA; Supplementary Fig. 7) and the PSI[10,15] indicated two alternative states in the overall microbiota composition, roughly corresponding to low and high *Prevotella*, in line with previous reports[9,10]. The main principal component explained 16% of the overall variation in the data and exhibited a continuous gradient associated with *Bacteroides* spp. and relatives of *Oscillospira guillermondii* and *Faecalibacterium prausnitzii*. The second principal component exhibited a bimodal distribution associated with the *Prevotella* (relatives of *P. oralis* and *P. melaninogenica)* group, explaining 13% of the total variation. We did not observe robust bimodality on the other major principal components. A complementary analysis using the principal coordinates analysis with UniFrac distances[28] yielded similar results, although the bimodality of the second (*Prevotella*) principal axis was less pronounced than in the standard PCA (Supplementary Fig. 7). Whereas the genus *Prevotella* and *Bacteroides* spp. have been earlier suggested as drivers of community-level enterotypes[9], our observations extend and complement this analysis by identifying specific bacteria exhibiting bistable properties within these groups and analysing their temporal dynamics, and by demonstrating that these groups in fact frequently co-occur in a host in varying combinations. The high-abundance state of the *Prevotella* group had the most dominating effect on the ecosystem, with an average relative abundance of 10% among the subjects with the high-abundance state (8 and 2% for the co-occurring *P. melaninogenica* and *P. oralis* groups, respectively). Due to their lower relative abundances (≤2% in the high-abundance state) compared with the *Prevotella* group, the bimodal patterns in the *B. fragilis*,

© 2014 Macmillan Publishers Limited. All rights reserved.

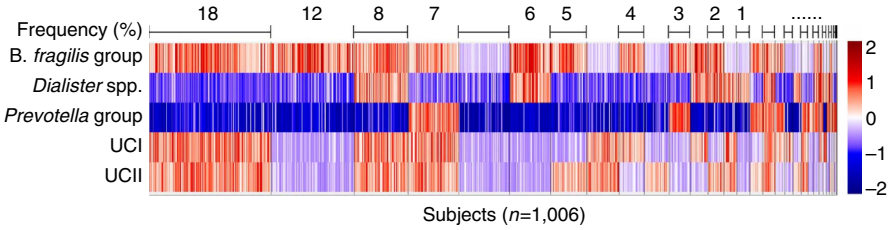


Figure 6 | The tipping elements co-occur in various combinations. The bimodal groups (rows) co-occur in various combinations within the study population. Shading indicates the relative abundance (log_{10}) for each group with respect to the identified tipping point between the alternative states of low (blue) and high (red) abundance. The 1,006 subjects are ordered based on the combination frequencies (top row). The most frequent combination (18%) corresponds to the high-abundance state of the UCI, UCII and *B. fragilis* groups combined with the low-abundance state of the *Dialister* spp. and *Prevotella* groups.

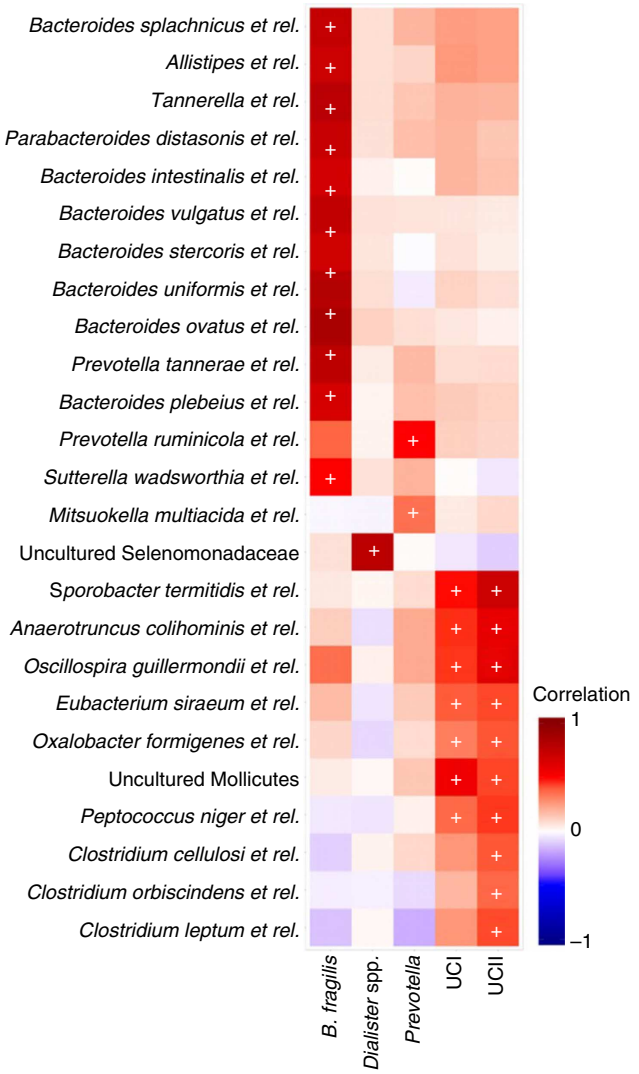


Figure 7 | Covariation between the bimodal groups and other taxa. Summary of correlations between the bimodal groups (columns) and the 124 other genus-like groups targeted by the HITChip microarray; significant correlations with 25 groups (rows; $P<0.05$; $|r|>0.25$; Pearson correlation) were consistently observed in independent sample sets analysed using two different DNA extraction methods (mechanical lysis; $n=401$ and enzymatic lysis-based method; $n=287$), and in 100 bootstrap data sets sampled from the overall data collection ($n=1,006$). The remaining 99 groups that did not show significant correlations have been removed for clarity. The 32 correlations that were found to be consistently significant in all tests are indicated by '+'. Shading indicates Pearson correlation between the taxa across the 1,006 western adults.

Dialister spp., UCI and UCII groups are more easily masked by variability in other, more smoothly varying taxa that correlate with these groups (Fig. 7), and hence the alternative states associated with these bacteria are easily overlooked in ecosystem-level analysis. Interestingly, whereas these less abundant groups had more moderate influence at the ecosystem-level variation, their high-abundance states were notably more prevalent compared with the high-abundance state of the more abundant *Prevotella* group, with the following population frequencies: *Prevotella* group (20%), *Dialister* spp. (28%), UCI (53%), UCII (57%) and *B. fragilis* (75%).

We also observed significant associations between the bistable bacteria, community diversity and the overall metagenomic gene richness (Wilcoxon test; $P<0.01$ for all groups, except *B. fragilis*; Supplementary Fig. 6; Supplementary Note 4) when we incorporated the recently reported metagenomic sequencing data that were available for 255 subjects[29] in our main cohort. In our analysis of these meta-genomes, the high-abundance state of the *B. fragilis* and *Dialister* spp. groups were found to be associated with a low gene count, in line with their reported association with pronounced metabolic dysfunction[29]. On the contrary, we observed associations between the high-abundance state of the *Prevotella*, UCI and UCII groups and the high gene count, which has been linked to a potentially healthy metabolic phenotype[29]. The increased metagenomic richness could be explained by the increased community diversity associated with these taxa; among the bistable groups, the UCI and UCII showed the strongest association with the overall community diversity ($\rho=0.5$; $P<0.001$; upper tail (AS89) test). Whereas the community diversity exhibited also mild positive correlation with the *Prevotella* group ($\rho=0.13$; $P=0.04$), the overall community diversity was higher in the subjects with high-abundance UCI/UCII than in the subjects associated with high-abundance *Prevotella* ($P<0.001$; *t*-test). The higher diversity associated with the UCI/UCII groups ($\leq 2\%$ relative abundance) could be partially explained by the more dominating status of the high-abundance *Prevotella* group (10% relative abundance), which tends to reduce the overall community diversity. We did not observe significant associations between the *B. fragilis* and *Dialister* spp. groups and community diversity ($|\rho|<0.07$; $P>0.3$).

Host factors and health status. The emerging picture is that rather than overarching alternative states corresponding to distinct community types, the natural variation of the human intestinal microbiota is reflected in specific bacterial groups representing 'tipping elements' that are only loosely coupled, specific components of the intestinal microbiota that exhibit alternative stable states linked to the overall microbiota composition and host factors. We have borrowed this analogy from the

© 2014 Macmillan Publishers Limited. All rights reserved.

climate system that is now also believed to have distinct tipping elements[11]. An important corollary is that the resilience (the capacity to recover from perturbations) of the alternative states will be influenced by a range of factors that may primarily affect specific subcommunities within the intestinal microbiota, rather than the overall ecosystem. Gradual changes in external or internal factors associated with the ecosystem changes may drive the implicated bistable bacteria towards a tipping point where an abrupt shift between the contrasting alternative states may follow[8,30]. Our analysis indicates that such shifts are often associated with changes in a wider set of covarying taxa (Fig. 7). One way to infer how resilience of the microbiota might depend on such factors is to use probability densities of bacterial abundance to estimate the basins of attraction. On the basis of this, we can observe, for instance, that a basin of attraction around an apparently resilient high-abundance state of UCI arises in late middle age (Fig. 8; Supplementary Table 3). This statistically significant association between UCI and age ($P<0.05$; χ^2 test) suggests that the resilience of the low-abundance state decreases with increasing age, and eventually, even a small perturbation may induce a shift to the high abundance state. Moreover, the analysis shows that population-level variation may mask alternative states that are more pronounced in specific subpopulations, such as particular age groups (Fig. 8b). According to a logistic regression model, each age decade increased the odds of the UCI high-abundance state by $2.6 \pm 1.2\%$ (95% confidence intervals), but the changes were particularly emphasized in the young and elderly populations. These findings are in line with previous studies that have reported notable changes in the overall microbiota composition in the elderly population[27,31,32], and with a recent study where differences at the community level were observed between young and elderly adults, but less during the middle age[31].

Overall, we observed a range of statistically significant associations between the bimodal taxa and host factors such as ageing, obesity and health status based on logistic regression ($P<0.05$; χ^2 test; see Methods). Among the 1,006 western adults, body mass index (BMI) was negatively associated with UCI and UCII. Age was positively associated with the UCI and UCII, and negatively with the *Dialister* spp. and *Prevotella* groups. Males had higher levels of bacteria belonging to the *Prevotella* group and lower levels of the *B. fragilis* group compared with women. Regarding nationality, the *Prevotella* group was underrepresented in Nordic and Southern European countries and UCII in South Europe, while the *Dialister* spp. and *B. fragilis* groups were particularly common in Northern Europe (Scandinavia, UK and Ireland). Analysis of an additional phylogenetic microarray data set derived from subjects with irritable bowel syndrome and severe obesity indicated association between these health complications and the low-abundance UCI and UCII (see Methods). Metabolic syndrome (MetS) was positively associated with *B. fragilis* (false discovery rate (FDR) $<5\%$), and showed a trend (FDR $<20\%$) towards positive association with *Dialister* spp. and a negative association with the *Prevotella* group (Supplementary Table 4). We did not observe associations between the bimodal groups and type II diabetes, ulcerative colitis or cardiovascular disease in this data collection. Overall, these results suggest that differences in health status can be reflected in changing state probabilities of specific taxa.

In conclusion, these observations provide substantial evidence for bistability in specific taxonomic groups within the human intestinal microbiota, including *Dialister* spp., the *Prevotella* group and the uncultured Clostridiales groups UCI and UCII.

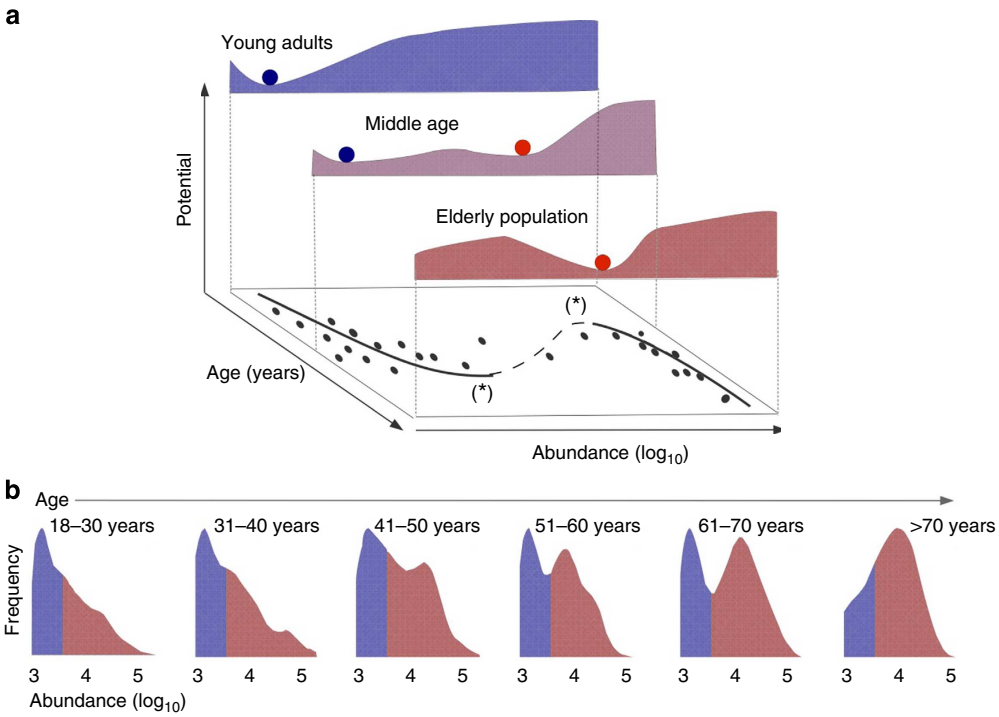


Figure 8 | Changing state probabilities associated with ageing. (**a**) A schematic illustration of shifting state probabilities in the UCI group associated with ageing. The UCI abundance data (bottom plane) suggest a catastrophe-fold, where the solid and dashed parts of the curve correspond to stable states and unstable equilibria, respectively. The depth and width of the potential minima indicate decreasing resilience of the system towards the bifurcation points (*) during ageing. (**b**) Observation density in various age groups highlights the associations between the UCI state probabilities and age (sample sizes: 220 (18–30 years); 147 (31–40 years); 192 (41–50 years); 258 (51–60 years); 114 (61–70 years); and 19 (71–77 years)). The alternative states of low (blue) and high (red) abundances are here more clearly pronounced than in the population-level histogram pooled over all age groups (Fig. 1).

© 2014 Macmillan Publishers Limited. All rights reserved.

These groups exhibited robust bimodality in their abundance distributions, and decreased temporal stability at the intermediate abundance range. A bimodal abundance distribution associated with the *B. fragilis* group was also observed, but we could not confirm the bistability of this group in subsequent analyses. Our findings were supported by independent subsets of the data using different DNA extraction methods and geographical distributions, and in additional data sets from dietary intervention trials, including 46 placebo-treated control subjects. Whereas the observed states were robust to short-term dietary differences, which could not explain the observed population-level bimodal patterns, this does not exclude the possibility that targeted dietary interventions could be potentially used to manipulate specific bimodal groups in a personalized manner; we did observe state shifts in individual subjects following dietary interventions, but these changes were masked by the overall within-state variation in the other subjects and did not reach statistical significance in our limited intervention data. The observed patterns were most pronounced in individual phylotypes and in specific subpopulations, such as different age groups, and linked to the overall ecosystem composition and host health.

Discussion

Our results demonstrate that despite the dominating nature of gradual variation in bacterial abundances in the intestinal ecosystem, specific bacterial groups form a limited number of contrasting, stable configurations that associate with host physiology and health. Bistability in bacterial abundances can be induced by many factors, such as disruptive selection[33], competition and cooperative feedback loops[34], or environmentally or genetically determined host factors. Hence, we believe that the here identified bistable taxa, together with other microbial co-occurrence relationships[26,35,36], will be instrumental in providing insight on the variation, regulation and health implications of the intestinal microbiota. Targeting specific subpopulations, as opposed to the daunting complexity and variability of the entire ecosystem, can simplify the characterization and possible manipulation of the intestinal microbiota. The implicated taxa may allow stratification of the subjects based on alternative states that in parallel to gene richness[29] represent promising novel targets for microbiome-based diagnostics and therapies[37]. Moreover, resetting these 'bacterial dual in-line package (DIP) switches' may be a radically new way to approach the rapidly growing number of health issues related to the intestinal microbiota, in analogy to the binary DIP switches from early computer era. Further intervention studies will be needed to address the causal relations with the overall community composition and host factors, and to investigate whether these alternative stable states translate into differential disease susceptibility or drug response of the host.

Methods

Sample collection. Phylogenetic microarray data from over 5,000 intestinal samples were available in the in-house MySQL database of the HITChip, a phylogenetic microarray[12]. To investigate the microbiota variability in western adult population, we selected from the HITChip database 1,006 western adults (18–77 years; average BMI 26.7 kg m^{-2}; s.d. 5.8 kg m^{-2}). To control the potentially confounding effect of other environmental or host factors, we curated the available sample annotations and excluded subjects who had reported diseases or antibiotic treatments, or were living outside western countries (Europe/the United States), and used bootstrap analysis to verify that the results were robust to random sampling fluctuations as described below. A single faecal sample per subject was selected; the first time point was used if multiple samples were available. The subjects cover 15 western countries (Europe and the United States) that were further aggregated into Central European (Belgium, Denmark, Germany, the Netherlands), Eastern European (Poland), Nordic (Finland, Norway, Sweden), South European (France, Italy, Serbia, Spain), UK/Ireland (UK, Ireland) and the United States (the United States) groups. To our knowledge, this is the largest collection of phylogenetic microarray profiling data of the human intestinal microbiota. Due to the retrospective nature of this study, we did not select a particular sample size based on power calculations or related approaches or replicate the experiments but instead included all available data, and performed various statistical tests and verifications in independent subsets of the data to assess the significance and robustness of the findings.

Microbiota profiling. Culture-independent techniques for characterizing intestinal microbiota include phylogenetic microarrays and sequencing-based approaches[38]. Phylogenetic microarrays can be used to quantify with high accuracy the known (previously detected) bacteria including the low-abundance phylotypes that often remain undetectable with conventional sequencing depths[39]. Our analysis is based on the phylogenetic HITChip microarray[12], produced by Agilent Technologies (Santa Clara, CA, USA), that provides a sensitive analysis platform to assess differences in relative abundance of intestinal bacteria. The HITChip microarray targets the V1 and V6 hypervariable regions of 16S rRNA gene, and can detect 1,033 species-like bacterial phylotypes (>98% sequence similarity in the 16S rRNA gene) that represent the majority of the bacterial diversity of the human intestine[12]. For the analyses, hybridization signals were summarized to 130 genus-like phylogenetic groups (>90% sequence similarity in the 16S rRNA gene) that are referred to as species and relatives, the latter being shortened as '*et rel.*'[12]. The HITChip microarray allows highly reproducible (98 ± 2% Pearson correlation across technical replicates) and deep analysis (reproducible detection of phylotype abundances below a 0.1% relative abundance level) of the phylotype composition of intestinal samples, comparable to 200,000 next-generation sequencing reads per sample[39]. Hence, the standardized analytical pipeline for the HITChip microarray data collection provides advantage over sequencing-based approaches, where the comparable collections of intestinal microbiota profiling data are smaller, and integration of data from individual studies is more challenging, for example, due to the use of different primers, continuously changing sequencing platforms and protocols. The faecal samples, collected in the context of clinical trials, were typically taken at home, frozen at −20 °C, delivered to the study centre within 24 h and stored at −80 °C. The DNA was extracted and prepared by amplification of the full-length 16S rRNA gene, followed by transcription and labelling of the resultant RNA with Cy3 and Cy5 before fragmentation and hybridization on the array[12]. The signal intensity data from the microarray hybridizations were collected using the G2505C scanner by Agilent and preprocessed using an in-house MySQL database and custom R scripts. Each scanner channel from the array was spatially normalized using polynomial regression, followed by outlier detection and filtering in each set of probes with a χ^2 test. The intensities were sample-wise quantile normalized and averaged to a single value for each probe–sample pair. Each sample was hybridized on at least two replicates to ensure reproducibility. Between-sample normalization was performed at probe level with the min–max normalization using 0.5% quantiles estimated from the data. Each targeted genus-like group is probed by multiple (median 19) oligos. Logarithmized ($\log_{10}$) signal of the probes targeting the same phylotype were summarized with Robust Probabilistic Averaging[40,41] (fRPA; RPA R package). The $\log_{10}$-transformed signals were used as a proxy for bacterial logarithmic abundance. As a proxy for relative abundance, we used the fraction of the total probe signal calculated for the probes targeting a given taxonomic group.

Bacterial abundance types. We categorized the 130 genus-like groups into characteristic 'abundance types' to gain an overview of their overall population-level variability. The abundance types included rare taxa and prevalent taxa with symmetric, (left- and right-) skewed and bimodal distributions (Fig. 2; Supplementary Table 1). The bimodal taxa were identified as described below, including verification in multiple data sets. For the other genus-like groups, the categorization was based on the 401 unique subjects whose faecal samples were processed with the repeated bead beating protocol that is based on mechanical cell disruption method for DNA extraction, and that we have found to be superior to alternative methods in terms of DNA extraction efficiency[25]. Altogether, 60 bacterial groups were abundant and prevalent, exceeding HITChip $\log_{10}$ signal of 3.0 in >20% of the subjects; the remaining 70 groups were considered rare. The 34 prevalent groups with symmetric log abundance (absolute skewness <0.5; e1071 CRAN R package) exhibited relatively constant logarithmic abundances across subjects, fluctuating symmetrically around a mean value. The skewed types included both rare and prevalent bacteria with both left- and right-skewed types (skewness < −0.5 and >0.5, respectively).

Potential analysis. We used potential analysis as the primary state identification method. This approach is based on the theory of stochastic dynamical systems, and was recently applied to identify alternative stable states in climatology and forest ecosystems[13,14]. In the following, we provide a brief summary of the underlying theory and assumptions.

Let us assume an underlying stochastic system that has a potential function of the form

$$dz = -U(z)dt + \sigma dW, \quad (1)$$

where $U(z)$ is the potential function, z is the state variable (in our case, the microbe

© 2014 Macmillan Publishers Limited. All rights reserved.

$\log_{10}$ abundance), σ is the noise level and dW is a noise term (Wiener process). The minima of the potential function correspond to stable states of the system. The corresponding Fokker–Planck equation connects the potential of this model to the probability density of the state variable z. The potential U is given by:

$$U = -\sigma^2 \log(p_d)/2 \quad (2)$$

where p_d is the empirically estimated probability density function of the state variable z. An example of the estimated probability density for the bimodal *Dialister* spp. in our data is shown in Supplementary Fig. 3. As our primary interest is in qualitative analysis of the potential function shape, we scaled the potential to the noise level (U/σ^2) (ref. 13). We estimated the probability density with Gaussian kernels (the R density function) and automated kernel width adjustment with Scott's method[42] (the *bw* function in the R stats package), giving the bandwidth $h = 1.06\, s/n^{1/5}$, where s is the s.d. of the data points and n is the sample size. We adjusted this bandwidth with a factor of 0.8 to achieve more sensitive detection of multimodality. We estimated the local minima and maxima numerically and reduced spurious findings by requiring that the difference between local minima and the closest maxima was larger than a threshold value[13] (0.005 in our analysis) and with a minimum density of 0.1. We used the number of local minima of the potential function to quantify the number of distinct modes[14].

Bimodality detection. We implemented the potential analysis as a state detection methodology[13,14] in the earlywarnings R package (http://www.early-warning-signals.org/; v. 1.0.59; livpotential_ews function). To avoid spurious findings associated with random sampling fluctuations and to quantify the robustness of bimodal patterns, we repeated this analysis in 100 bootstrap data sets, and selected for further analysis the taxonomic groups that exhibited consistent evidence for multimodality in $\geq$90% of the bootstrap data sets (Supplementary Note 5). While the estimated number of modes in some cases varied between the bootstrap data sets, the evidence was the strongest for two alternative states in all multimodal groups. Hence, we used the overall fraction of bootstrap samples that indicated deviation from unimodality to quantify the evidence for bimodality. Our sample collection included samples from multiple studies extracted using different DNA extraction methods, which could affect bimodality detection in a taxon-specific manner[25]. To control this, we confirmed the bimodality of the reported six bimodal groups in the subset of 401 samples extracted using the mechanical lysis-based DNA extraction method (repeated bead beating)[25]. We also assessed bimodality in the subset of 287 samples processed using enzymatic lysis. The UCII group had a moderate (83%) bootstrap support for bimodality, whereas bimodality of the *B. fragilis* group was not supported in this smaller data set (31% bootstrap support). The bimodal patterns in *Dialister* spp., *P. melaninogenica*, *P. oralis* and UCI were verified in 100% of the bootstrap samples also in this data set. For the final analysis, we included the *B. fragilis*, *Dialister* spp., *P. melaninogenica*, *P. oralis*, UCI and UCII groups whose bimodality was detected in the overall data collection and in at least one of the two distinct DNA extraction method subsets with >90% bootstrap support (Supplementary Note 6). An independent analysis of bimodality in these univariate observations of bacterial abundance was performed using the PSI (with $M = 20$ random splits into training and test sets; fpc R package) based on the partition around medoids clustering and Euclidean distance[10,15].

Tipping point. The tipping point is here defined as the particular bacterial abundance value that maximizes the potential function of a bistable system; this corresponds to an unstable threshold between the contrasting alternative stable states, which correspond to the local minima of the potential function. The potential minima and maxima were determined by potential analysis. We used the average over the different bootstrap data sets as the final tipping point estimate; this provides increased robustness against random sampling fluctuations.

Temporal analysis. To investigate the temporal stability of the observed alternative states, we analysed 78 subjects from the main cohort with additional follow-up data from dietary intervention trials; 28% of the follow-up samples were subject to dietary intervention but no significant changes following the dietary intervention were seen in the bimodal taxa, except for the UCI and UCII groups ($P < 0.05$; t-test). The median age of the subjects was 49 years (27–63 years); the median BMI 33 kg m^{-2} (18–43 kg m^{-2}); most subjects ($n = 65$; 86%) were females from Finland ($n = 43$; 55%) and Belgium ($n = 32$; 41%); and 81% of the subjects had 2–3 time points; the remaining 19% had 4–5 time points. Most (95%) time points were within 1–9 months from the baseline (mean 3.5 months), with the average follow-up interval (time between consecutive time points) of 2.4 months. Despite some variability between subjects, the follow-up times were within a similar time span and the differences in the follow-up times did not explain differences in the observed fluctuations between time points, as these two variables were not significantly correlated in our data ($|\rho| < 0.05$; $P > 0.05$; upper tail (AS89) test). We quantified the stability at the intermediate abundance range by calculating the correlation between the distance from the intermediate (50%) abundance quantile and the observed shift between the consecutive time points; negative correlations indicate instable regions where larger shifts in microbial abundance are observed on average. In addition, we approximated the state shift frequencies in subjects with varying follow-up times based on Kaplan–Meier survival analysis (the survival R package), which gives an estimate of the fraction of subjects remaining in their initial state as a function of time based on limited observations with varying follow-up times. We used the estimated fraction of subjects remaining in their original state at the 3-month time point to quantify state stability (Supplementary Note 7). Since the observed state shifts could arise from natural, continuous fluctuations in bacterial abundance rather than abrupt shifts between alternative stable states, we used the same procedure to estimate how often state shifts would be observed in the other prevalent taxa based on simulated states of low and high abundance based on the same abundance quantiles to determine the 'tipping point' between the two states, and compared the bistable groups with the other prevalent taxa (Supplementary Fig. 2b).

Community-level analysis. We assessed the evidence for alternative states at the overall ecosystem level based on PCA (Supplementary Fig. 7a) and the partition around medoids multivariate clustering with Jensen–Shannon dissimilarity, coupled with the PSI to determine the optimal cluster number as this combination was earlier shown to have a good overall performance[10] (Supplementary Note 8). In addition, we used principal coordinates analysis (that is, classical multidimensional scaling; cmdscale function of the R stats package)[43] as a complementary nonlinear approach for community-level visualization[10] based on generalized UniFrac distances[28] (GUniFrac function of the R GUniFrac package[44]) that take into account the phylogenetic structure of the community, quantified by HITChip probe-level analysis (Supplementary Fig. 7b).

Short-term dietary interventions. Whereas dietary information was not available for our main data set, we analysed the effects of short-term dietary interventions in an additional data set from the HITChip database consisting of 138 subjects sampled during dietary intervention studies, including three previously published[16,17,45] and two unpublished dietary intervention trials: (i) 52 Finnish adults fulfilling the criteria for MetS that were subjected to qualitative changes in carbohydrate intake[16] (50% of the samples in a control group); (ii) 33 obese Belgian women from a prebiotic (inulin-type fructan) intervention study[45] (50% of the samples in a control group); (iii) 31 obese Finnish subjects with weight-loss diet (HITChip project 'ISRCTN67529475'); (iv) 8 Norwegian subjects with a vegan diet (HITChip project 'Vegan'); and (v) 14 obese British males with fibre-supplemented and weight-loss diets[17]. In addition to the dietary intervention subjects, this data collection included 46 placebo-treated control subjects.

Associations with host parameters. We quantified associations between the bimodal taxa and host factors by using subject metadata (age, BMI, sex, nationality) and DNA extraction method as fixed effects in multiple logistic regression to predict the state of the corresponding bacteria in each subject. The significance of each parameter was estimated by log-ratio (χ^2) test comparing the models with and without the corresponding factor in predicting the state (the GLM function in the R stats package). Missing annotations were omitted, and the FDR was estimated to correct for multiple testing based on the Benjamini–Hochberg method with the p.adjust function (R stats package).

Health associations. To investigate the associations between the alternative states and health status, we compared the 1,006 adults with no reported health complications with additional samples in the HITChip database representing western adults with irritable bowel syndrome ($n = 106$), MetS ($n = 66$), type II diabetes ($n = 78$), ulcerative colitis ($n = 62$) and cardiovascular disease ($n = 45$). In addition, we divided the non-compromised subjects into severe obese (BMI$\geq$35; $n = 136$) and other (BMI$<$35; $n = 870$) subjects. We used multiple logistic regression to predict disease occurrence from logarithmic abundance of each bimodal bacterial group. The DNA extraction method, age, BMI and sex were included in the model as potentially confounding variables. The P values from log-ratio test were corrected using the Benjamini–Hochberg procedure. For exploratory purposes, associations with FDR$<$20% were considered to establish a trend and FDR$<$5% to be significant (Supplementary Table 4). We compared the performance of the logistic regression model with and without bacterial abundance based on receiver operator characteristic analysis to detect disease occurrence; the average area under curve (AUC) values were derived from fivefold cross-validation. While our analysis suggests significant association between MetS and specific taxa, the differences in the AUC values between the models with and without bacterial abundance were small (0.1 increase with bacterial abundance) in the MetS group, indicating limited predictive ability in our data to distinguish the effects of MetS from the effects of age, sex and BMI. In severe obesity, inclusion of bacterial abundance notably improved disease prediction (AUC 0.58 versus 0.69 for UCI; 0.60 versus 0.68 for UCII).

References

1. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464,** 59–65 (2010).
2. Zoetendal, E. G., Rajilić-Stojanović, M. & de Vos, W. M. High-throughput diversity and functionality analysis of the gastrointestinal tract microbiota. *Gut* **57,** 1605–1615 (2008).

© 2014 Macmillan Publishers Limited. All rights reserved.

3. de Vos, W. M. & de Vos, E. A. J. Role of the intestinal microbiome in health and disease: from correlation to causation. *Nutr. Rev.* **70,** S45–S56 (2012).
4. Cho, I. & Blaser, M. J. The human microbiome: at the interface of health and disease. *Nat. Rev. Genet.* **13,** 260–270 (2012).
5. Costello, E. K., Stagaman, K., Dethlefsen, L., Bohannan, B. J. M. & Relman, D. A. The application of ecological theory toward an understanding of the human microbiome. *Science* **336,** 1255–1262 (2012).
6. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486,** 207–214 (2012).
7. Jeffery, I. B., Claesson, M. J., O'Toole, P. W. & Shanahan, F. Categorization of the gut microbiota: enterotypes or gradients? *Nat. Rev. Microbiol.* **10,** 591–592 (2012).
8. Scheffer, M. *Critical Transitions in Nature and Society* (Princeton Univ. Press, 2009).
9. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473,** 174–180 (2011).
10. Koren, O. *et al.* A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. *PLoS Comput. Biol.* **9,** e1006863 (2013).
11. Lenton, T. M. *et al.* Tipping elements in the earth's climate system. *Proc. Natl Acad. Sci. USA* **105,** 1786–1793 (2008).
12. Rajilić-Stojanović, M. *et al.* Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. *Environ. Microbiol.* **11,** 1736–1751 (2009).
13. Hirota, M., Holmgren, M., van Nes, E. H. & Scheffer, M. Global resilience of tropical forest and savanna to critical transitions. *Science* **334,** 232–235 (2011).
14. Livina, V. N., Kwasniok, F. & Lenton, T. M. Potential Analysis reveals changing number of climate states during the last 60 kyr. *Clim. Past* **6,** 77–82 (2010).
15. Tibshirani, R. & Walter, G. Cluster validation by prediction strength. *J. Comput. Graph. Stat.* **14,** 511–528 (2005).
16. Lappi, J. *et al.* Intake of whole-grain and fiber-rich rye bread versus refined wheat bread does not differentiate intestinal microbiota composition in Finnish adults with metabolic syndrome. *J. Nutr.* **143,** 648–655 (2013).
17. Salonen, A. *et al.* Impact of diet and individual variation on intestinal microbiota composition and fermentation products in obese men. *ISME J.* doi:10.1038/ismej.2014.63 (2014).
18. David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **23,** 559–563 (2014).
19. Martínez, I. *et al.* Gut microbiome composition is linked to whole grain-induced immunological improvements. *ISME J.* **7,** 269–280 (2013).
20. Kato, H. *et al.* Colonisation and transmission of Clostridium difficile in healthy individuals examined by PCR ribotyping and pulsed-field gel electrophoresis. *J. Med. Microbiol.* **50,** 720–727 (2001).
21. van Nood, E. *et al.* Duodenal infusion of donor feces for recurrent *Clostridium* difficile. *N. Engl. J. Med.* **368,** 407–415 (2013).
22. Gough, E., Shaikh, H. & Manges, A. R. Systematic review of intestinal microbiota transplantation (fecal bacteriotherapy) for recurrent *Clostridium* difficile infection. *Clin. Infect. Dis.* **53,** 994–1002 (2011).
23. Scheffer, M. *et al.* Early-warning signals for critical transitions. *Nature* **461,** 53–59 (2009).
24. Dridi, B., Henry, M., El Khéchine, A., Raoult, D. & Drancourt, M. High prevalence of Methanobrevibacter smithii and Methanosphaera stadtmanae detected in the human gut using an improved DNA detection protocol. *PLoS One* **4,** e7063 (2009).
25. Salonen, A. *et al.* Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: effective recovery of bacterial and archaeal DNA using mechanical cell lysis. *J. Microbiol. Methods* **81,** 127–134 (2010).
26. Faust, K. *et al.* Microbial co-occurrence relationships in the human microbiome. *PLoS Comput. Biol.* **8,** e1006606 (2012).
27. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486,** 222–227 (2012).
28. Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* **71,** 8228–8235 (2005).
29. Chatelier, E. L. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500,** 541–546 (2013).
30. Scheffer, M., Carpenter, S., Foley, J. A., Folke, C. & Walker, B. Catastrophic shifts in ecosystems. *Nature* **413,** 591–596 (2001).
31. Lan, Y., Kriete, A. & Rosen, G. L. Selecting age-related functional characteristics in the human gut microbiome. *Microbiome* **1,** 2 (2013).
32. Biagi, E. *et al.* Through ageing, and beyond: gut microbiota and inflammatory status in seniors and centenarians. *PLoS ONE* **5,** e10667 (2010).
33. Rueffler, C., Van Dooren, T. J. M., Leimar, O. & Abrams, P. A. Disruptive selection and then what? *Trends Ecol. Evol.* **21,** 238–245 (2006).
34. Faust, K. & Raes, J. Microbial interactions: from networks to models. *Nat. Rev. Microbiol.* **10,** 538–550 (2012).
35. Claesson, M. J. *et al.* Gut microbiota composition correlates with diet and health in the elderly. *Nature* **488,** 178–184 (2012).
36. Holmes, I., Harris, K. & Quince, C. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS ONE* **7,** e30126 (2012).
37. Lemon, K. P., Armitage, G. C., Relman, D. A. & Fischbach, M. A. Microbiota-targeted therapies: an ecological perspective. *Sci. Transl. Med.* **4,** 137rv5 (2012).
38. Kuczynski, J. *et al.* Experimental and analytical tools for studying the human microbiome. *Nat. Rev. Genet.* **13,** 47–58 (2011).
39. Claesson, M. J. *et al.* Comparative analysis of pyrosequencing and a phylogenetic microarray for exploring microbial community structures in the human distal intestine. *PLoS ONE* **4,** e6669 (2009).
40. Lahti, L., Torrente, A., Elo, L. L., Brazma, A. & Rung, J. A fully scalable online pre-processing algorithm for short oligonucleotide microarray atlases. *Nucleic Acids Res.* **41,** e110 (2013).
41. Lahti, L., Elo, L. L., Aittokallio, T. & Kaski, S. Probabilistic analysis of probe reliability in differential gene expression studies with short oligonucleotide arrays. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **8,** 217–225 (2011).
42. Scott, D. W. *Multivariate Density Estimation: Theory, Practice, and Visualization* (John Wiley & Sons, 1992).
43. Gower, J. C. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53,** 325–328 (1966).
44. Chen, J. *et al.* Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics* **28,** 2106–2113 (2012).
45. Dewulf, E. M. *et al.* Insight into the prebiotic concept: lessons from an exploratory, double blind intervention study with inulin-type fructans in obese women. *Gut* **62,** 1112–1121 (2013).

Acknowledgements

We thank Outi Immonen, Philippe Puylaert and Jarmo Ritari for technical support with the HITChip data analysis, and Vasilis Dakos and Erwin Zoetendal for their feedback. This work was supported by Academy of Finland (256950 to L.L.; 141140 to W.M.d.V.); European Research Council (ERC250172 to W.M.d.V.; ERC268732 to M.S.); Spinoza Awards and Gravitation Programmes, Netherlands Organization for Scientific Research (2008/2014 to W.M.d.V.; 2009/2014 to M.S.); and Alfred Kordelin foundation (to L.L.).

Author contributions

L.L., M.S. and W.M.d.V. designed the study. L.L. carried out data processing and statistical analysis, coordinated the study and prepared the manuscript. A.S., L.L., J.S. and W.M.d.V. contributed to acquisition of data. J.S. contributed to data analysis. All authors contributed to interpreting the results and writing the manuscript.

Additional information

Accession codes: The phylogenetic HITChip microarray data matrix and anonymized subject metadata have been deposited in the Dryad Digital Repository: http://doi.org/10.5061/dryad.pk75d

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Lahti, L. *et al.* Tipping elements in the human intestinal ecosystem. *Nat. Commun.* 5:4344 doi: 10.1038/ncomms5344 (2014).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

© 2014 Macmillan Publishers Limited. All rights reserved.

Towards a systems view of IBS

Emeran A. Mayer, Jennifer S. Labus, Kirsten Tillisch, Steven W. Cole and Pierre Baldi

Abstract | Despite an extensive body of reported information about peripheral and central mechanisms involved in the pathophysiology of IBS symptoms, no comprehensive disease model has emerged that would guide the development of novel, effective therapies. In this Review, we will first describe novel insights into some key components of brain–gut interactions, starting with the emerging findings of distinct functional and structural brain signatures of IBS. We will then point out emerging correlations between these brain networks and genomic, gastrointestinal, immune and gut-microbiome-related parameters. We will incorporate this new information, as well as the reported extensive literature on various peripheral mechanisms, into a systems-based disease model of IBS, and discuss the implications of such a model for improved understanding of the disorder, and for the development of more-effective treatment approaches in the future.

Mayer, E. A. *et al. Nat. Rev. Gastroenterol. Hepatol.* **12**, 592–605 (2015); published online 25 August 2015; doi:10.1038/nrgastro.2015.121

Introduction

IBS is the most common functional gastrointestinal disorder, occurring in up to 15% of the population worldwide.[1] Even though the syndrome is defined by chronically recurring abdominal pain and discomfort associated with altered bowel habits in the absence of organic disease, increased trait anxiety, as well as comorbidity with psychiatric and other chronic pain syndromes are common.[2] Despie an extensive body of reported information about peripheral[3–5] and central[6–9] mechanisms involved in the pathophysiology of IBS symptoms, and the development of animal models with high face and construct validity (that is, the models have many features similar to and is based on a similar pathophysiological concept as the human disease),[10] no comprehensive disease model has emerged that would guide the development of novel and effective therapies. This aspect is surprising in view of the comprehensive data as well as clinical experience demonstrating the strong relationship between psychosocial factors and IBS symptoms,[11] and in view of breakthroughs in the identification of several IBS-related biological abnormalities at several levels: the gut epithelium;[12] immune system;[4,5] neuroendocrine mechanisms;[13] brain structure and function;[7] stress response;[14] affective,[8,15] cognitive[6,16–18] and pain modulation[19,20] abnormalities; gene polymorphisms;[3] and the gut microbiome.[21,22] Despite these insights, a comprehensive understanding of how these various factors interact, and particularly to what degree they are involved, in the generation of symptoms in IBS in general or in IBS subsets (as opposed to representing epiphenomena) has not emerged. The long history of IBS research[23] is full of examples of reported abnormalities (including excessive mucus production, alterations in sigmoid colon motility, slow wave abnormalities in smooth muscle, gut inflammation and others), which were reported in small sample cohorts and often not confirmed in validation sample cohorts. Typically, the majority of reported mean differences are small when compared with healthy individuals, often do not take into account sex-related differences, might only be present in subsets of patients and correlations with clinical symptoms are weak. Thus, only a very small number of findings have translated into highly effective therapies.[24] These therapies include the serotonin (5-HT) 5-HT_3 receptor antagonist, alosetron, the 5-HT_4 agonist, tegaserod, the guanylate cyclase agonist, linaclotide, and the chloride channel blocker lubiprostone (for chronic idiopathic constipation).[25] Even though effective, the first two agents were restricted or withdrawn due to serious adverse effects, whereas the latter two are primarily targeted at treating the symptom of constipation.[25]

The brain and gut show reciprocal interactions in health and disease. For example, altered brain outputs via the autonomic nervous system (ANS) and the hypothalamic–pituitary–adrenal (HPA) axis have been shown to influence intestinal motility and secretion,[26] intestinal epithelial permeability,[3,12,27] immune function[28] and gut microbial composition.[21] In addition to these well-established influences of the brain on peripheral target cells, peripheral environmental gut-directed factors—in particular, dietary factors[29] and intestinal pathogens[30]—might have an equally important role on these same processes. Regardless of primary cause for the observed peripheral findings, several of them (in particular immune and microbiota-related signalling) can feed back to the brain, setting up circular regulatory loops (Figure 1).

Major barriers exist for progress towards a comprehensive understanding of IBS pathophysiology, which incorporates such circular regulatory loops. Despite a wealth of information supporting all mechanisms listed already, controversy regarding the primary role of the brain versus peripheral factors has persisted in the field. The majority

Departments of Medicine, University of California at Los Angeles, 10833 Le Conte Avenue, Los Angeles, CA 90095-7378, USA (**E.A.M.**, **J.S.L.**, **K.T.**, **S.W.C.**). West Los Angeles VA Medical Center, 11301 Wilshire Boulevard, Los Angeles, CA 90073, USA (**K.T.**). Institute for Genomics and Bioinformatics, University of California at Irvine, 4038 Bren Hall, Irvine, CA 92697-3435, USA (**P.B.**).

Correspondence to: E.A.M.
emayer@ucla.edu

Competing interests
The authors declare no competing interests.

© 2015 Macmillan Publishers Limited. All rights reserved

Key points

- Physiological and molecular alterations have been identified in the brain–gut axis of human and rodent models of IBS, yet a comprehensive disease model to guide effective drug development has not emerged
- Studies have identified distinct brain signatures in patients with IBS, which provide plausible neurobiological substrates of many previously reported behavioural and psychosocial observations
- Emerging evidence demonstrates correlations of these brain signatures with alterations in genetics, immune system and gut microbiota in IBS, even though the causality of these interactions remains unknown
- A systems-biology-based model is proposed to integrate the growing number of central, peripheral and behavioural IBS-related alterations, and to identify targets for more effective therapies

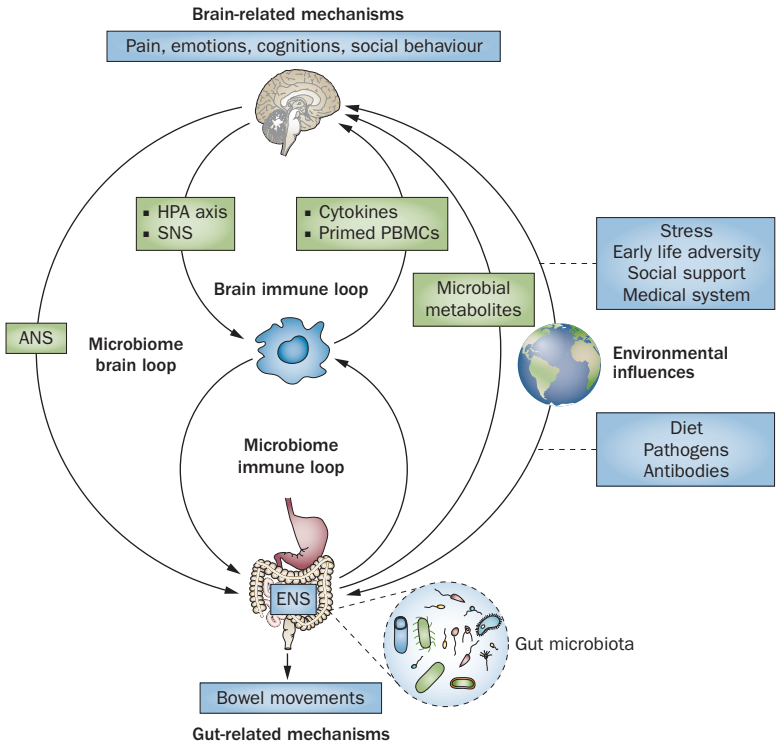


Figure 1 | Brain–gut axis. Schematic of the brain–gut axis, including inputs from the gut microbiota, the ENS, the immune system and the external environment. The model includes both peripheral and central components, which are in bidirectional interactions. Bottom-up influences are shown on the right side, top-down influences on the left side of the graph. Abbreviations: ENS, enteric nervous system; HPA, hypothalamic–pituitary–adrenal; PBMC, peripheral blood mononuclear cell; SNS, sympathetic nervous system. Modified with permission from Nature Publishing Group © Irwin, M.R. & Cole, S.W. *Nat. Rev. Immunol.* **11**, 625–632 (2011).[103]

of research and drug development have focused on single, usually peripheral, targets in preclinical models (such as ion channels or specific receptors on individual neurons). Although such studies have provided major insights into mechanisms of visceral pain,[10] their relevance for IBS symptom generation in humans has not firmly been established. Integration of multiple clinical, psychosocial and biological (genetic, immune, neurobiological) findings into a comprehensive disease model has yet to be achieved. For example, few studies aimed at psychosocial aspects of IBS have taken neurobiological concepts into account. The focus has been on descriptive, symptom-based rather than biology-based disease definitions. For example, different biological mechanisms might be associated with similar clinical presentations. The comprehensive identification of distinct biology-based subgroups of patients (including those based on sex), with different underlying pathophysiological components, who are differentially responsive to specific therapies has also not been achieved. A good example illustrating this point is a report on subsets of patients with IBS based on gut microbial signatures.[31]

In this Review, we will first describe novel insights into some key components of brain–gut–microbiome axis, starting with evolving concepts about alterations in defined structural and functional brain networks. We will discuss emerging evidence on how these brain network alterations are correlated with the immune system and the gut microbiota. We will then incorporate this information, as well as the reported extensive literature on various peripheral mechanisms, into a systems-based disease model of IBS. Rather than simply synthesizing the current knowledge about the disease, we will use this model to discuss the implications for better understanding of the disorder and for the development of more-effective treatment approaches in the future.

The nervous system component

Several clinical observations support a major role of the brain in IBS symptoms: the brain is ultimately responsible for generating the subjective experience of abdominal pain, discomfort and anxiety; stressful life events in early life have a major role in vulnerability to develop IBS, and psychosocial stressors in adulthood play crucial parts in the first onset, and perceived severity of symptoms;[32] centrally targeted pharmacological treatments and cognitive behavioural strategies are some of the most-effective treatment strategies.[33] The role of the enteric nervous system (ENS) in the regulation and coordination of motor and secretory functions of the gastrointestinal tract, in sensory function and its interactions with the brain have been extensively reviewed.[34,35] Consistent with the theme of this Review, the complexity of the interactions between multiple gut-based cell types (intrinsic and extrinsic sensory neurons, enteric glia, immune cells and innervated enteroendocrine cells) has been referred to as the 'gut connectome'.[36] The interactions between the ENS and central nervous system (CNS), two key components of the brain–gut axis, are mediated by the sympathetic and parasympathetic branches of the ANS,[6] and by multiple sensory and endocrine pathways.[34] Although it is difficult to characterize alterations of the ENS or of ANS-mediated brain–gut connections in living humans directly, functional, structural and metabolic brain imaging approaches have become a highly productive avenue to study brain–gut–microbiota interactions in health and disease.[8,9,37,38]

Although it has long been assumed that specific brain functions such as pain processing, emotion or cognition are attributable to the isolated operations of single brain regions, these processes are now viewed as resulting from the dynamic interactions of distributed brain areas operating in large-scale networks (Figure 2). These networks and their properties have been assessed by

© 2015 Macmillan Publishers Limited. All rights reserved

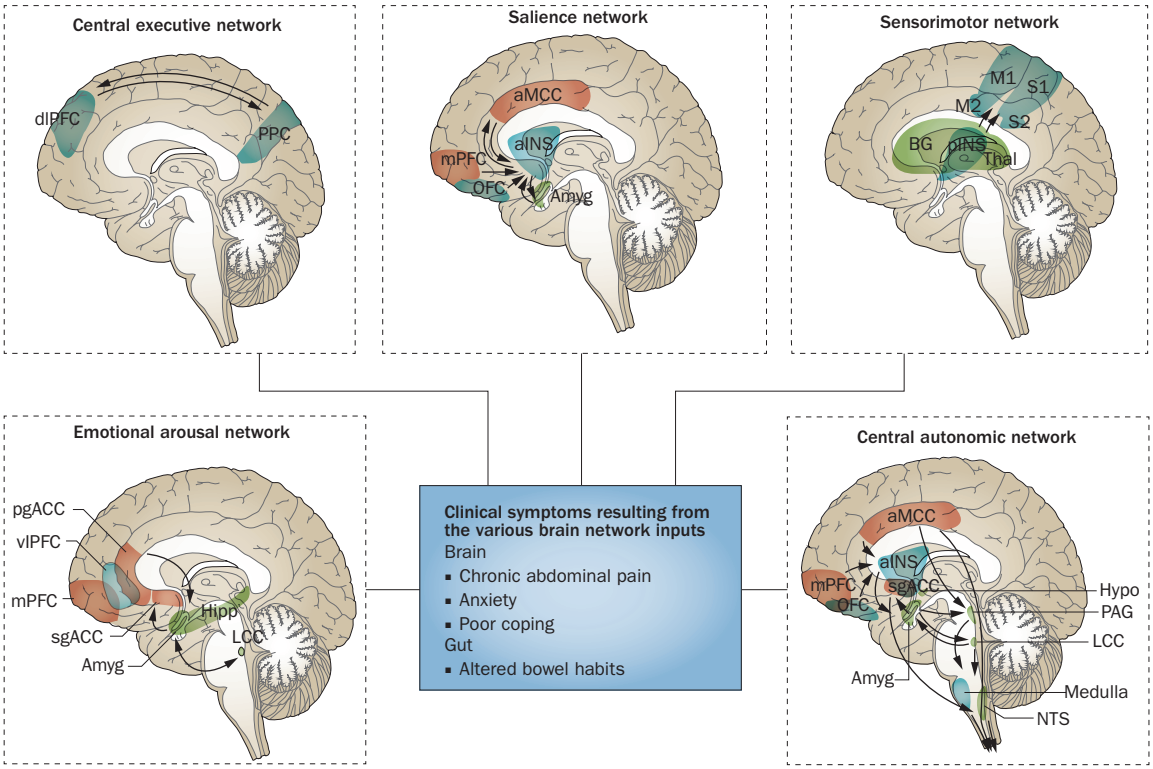


Figure 2 | Brain networks contributing to IBS symptoms. Depicted are task-related brain networks that have been described in the literature and for which structural and functional alterations have been reported in patients wiht IBS. The box in the centre describes the clinical symptoms related to the network inputs. Outputs most relevant for IBS pathophysiology occur in the form of descending pain modulation and autonomic nervous system activity. Abbreviations: Amyg, amygdala; aINS, anterior insula; aMCC, anterior midcingulate cortex; BG, basal ganglia; dlPFC, dorsolateral prefrontal cortex; Hipp, hippocampus; Hypo, hypothalamus; LCC, locus coeruleus complex; M1, primary motor cortex; M2, supplementary motor cortex; mPFC, medial prefrontal cortex; NTS, solitary nucleus; OFC, orbitofrontal cortex; PAG, periaqueductal grey; pgACC, pregenual anterior cingulate cortex; pINS, posteria insula; PPC, posterior parietal cortex; sgACC, subgenual anterior cingulate cortex; Thal, thalamus; vlPFC, ventrolateral prefrontal cortex.

using neuroanatomical and neurophysiological studies in animals,[10] as well as different brain imaging techniques and analyses in humans.[39–46] In humans, several types of networks have been reported: functional brain networks based on evoked responses[37] or intrinsic connectivity of the brain during rest;[39–41,44,45] structural networks based on grey matter parameters[47] and white matter properties; and anatomical networks based on white matter connectivity.[48] Both evoked and resting state studies performed in patients with IBS have demonstrated abnormalities in regions and task-related networks linked to emotional arousal,[49–53] central autonomic control,[6,54–56] central executive control,[51,57,58] sensorimotor processing[6,59–61] and salience detection.[57,62] IBS-related alterations in these networks have provided plausible neurobiological substrates for several information processing abnormalities reported in patients with IBS, such as biased threat appraisal and expectancy of outcomes (for example, salience network), autonomic hyperarousal (emotional arousal and central autonomic networks), and symptom-focused attention (central executive network).[63]

In the next section, we will discuss IBS-related changes that have been identified in these task-related networks. The individual networks are depicted in Figure 2, and their basic properties are described in Table 1.

Emotional arousal network

This network acts as an important link between stimulus appraisal (salience network) and ANS output (generated in the central autonomic network) to peripheral targets (gastrointestinal tract, gut microbes, immune system), thereby having an important role in determining the magnitude and duration of autonomic modulation of various gut functions. A reduction in the inhibitory feedback loop within the arousal network has been demonstrated in patients with IBS compared with healthy individuals,[51,64,65] and in healthy individuals as controls after decreasing central serotonin levels by acute tryptophan depletion.[66] Several studies published during the past decade support an increased responsiveness of emotional arousal circuits in relation to both expected and to delivered visceral stimuli, particularly in women.[67–76] A meta-analyses of functional MRI studies published between 2000 and 2010 demonstrated that during controlled rectal distension, patients with IBS show more consistent activation in regions associated with emotional arousal than healthy individuals.[53] Emotional arousal circuit reactivity is associated with 5-HT-related gene polymorphisms.[77] IBS-related functional alterations are accompanied by structural brain alterations in key regions of this network.[78]

© 2015 Macmillan Publishers Limited. All rights reserved

Table 1 | Brain networks with relevance to IBS

Network	Core regions and inputs	Key features
Default mode network	Medial prefrontal cortex, posterior cingulate or retrosplenial cortex, inferior parietal cortex, lateral temporal cortex and hippocampal formation[179,180]	Comprised of brain regions whose activity is greater during rest than goal-directed task performance[181] Associated with self-related processing, including monitoring internal thoughts and future planning[179,182,183]
Emotional arousal network	Amygdala, its positive connections with the locus coeruleus complex and inhibitory feedback projections to the amygdala from prefrontal and anterior cingulate subregions, and from the hippocampus	Activated by perceived or real perturbation of the organism's homeostasis Generates rapid feedback inhibition of amygdala,[184–186] thereby limiting the magnitude and duration of network activity and related activity in the central autonomic network
Central autonomic network	Control centres in the pontine-medulla (including PAG and hypothalamus), the central nucleus of the amygdala, and several cortical regions (including the anterior INS, ACC, prefrontal and motor regions)[187,188]	Regions related to SNS control overlap with the executive-processing and salience-processing networks (including the ventral anterior INS), whereas regions related to parasympathetic control are more associated with the default mode network[154] Provides central control and modulation of the ANS Involved in regulating respiratory, cardiovascular, endocrine and digestive system activity during cognitive, affective and motor tasks and sensation
Sensorimotor network	Core cortical regions are primary somatosensory cortex (S1, post central gyrus), primary motor cortex (M1; precentral gyrus), secondary somatosensory cortex (S2) and supplemental motor area[189] Close connections exist between the posterior INS (primary interoceptive cortex) and S1 Sensorimotor network connectivity to the thalamus, which relays peripheral sensory information to the cortex,[190] is established by 2 years of age[191]	Receives sensory input from the periphery and is important for awareness of body sensations and generation of appropriate motor responses Primary and secondary motor cortex, through their projections to the central autonomic network, might have a modulatory role in the sympathetic control of visceral function[192]
Central executive network	Lateral prefrontal cortices and posterior parietal cortex[193]	Activated during tasks involving executive functions such as attention, working memory, planning and response selection Often coactivated with regions of the salience network,[39] as the brain attempts to focus its limited processing capacity to only salient information via attention, working memory, planning and response selection[193]
Salience network	Dorsal ACC and anterior INS Core regions have strong connections to medial prefrontal and temporal regions, and subcortical regions including the amygdala, PAG and basal ganglia[194–196] Dorsal portions of the anterior INS receives prefrontal input, whilst ventral portions are closely linked with the amygdala and emotional arousal system	Anterior INS can be considered the main hub in the brain, switching from default mode network to activity-related networks, and coordinating and adjusting bodily and behavioural responses to environmental changes Responds to subjective salience of any interoceptive and exteroceptive stimulus reaching the brain, or to the expectation of such stimuli, and coordinates appropriate attentional, behavioural, affective and visceral responses to such stimuli[39,197] Responds with the most appropriate responses to biologically and cognitively relevant stimuli based on maintaining homeostasis[193,194] regardless of whether the subject is awake, engaged in a particular task, or asleep (achieved through close salience network connections with the other networks)

Abbreviations: ACC, anterior cingulate cortex; ANS, autonomic nervous system; INS, insula; PAG, periaqueductal grey; SNS, sympathetic nervous system.

Central autonomic network

An extensive body of knowledge derived from rodent studies[26,79–82] supports an upregulation of brain circuits in response to chronic stress, in particular the input from the locus coeruleus complex to the amygdala and the hypothalamus. Information from human brain imaging studies has been more limited due to the technical difficulties in studying key regions of this network, such as the hypothalamus, locus coeruleus complex and subnuclei of the amygdala, primarily due to limited spatial resolution. However, pharmacological brain imaging studies have implicated alterations in the corticotropin-releasing factor (also known as corticoliberin)–corticotropin-releasing receptor 1[64,83] and norepinephrine–α adrenergic receptor signalling system in this network.[84]

Sensorimotor network

As with other chronic pain disorders,[85–88] evidence indicates that IBS might be associated with alterations in brain networks concerned with the central processing and modulation of viscerosensory and somatosensory information. For example, compared with healthy individuals, patients with IBS showed increased low frequency power of spontaneous brain oscillations (suggesting increased

© 2015 Macmillan Publishers Limited. All rights reserved

neural activity) in regions belonging to the sensorimotor network.[59] These functional changes seem to be accompanied by structural changes in white and grey matter. For example, patients with IBS had widespread microstructural white matter abnormalities in sensory processing and/or modulation regions.[89] Female patients with IBS have cortical thickness increases in sensorimotor areas that correlated with clinical measures of symptom severity compared with healthy individuals.[90] Such grey matter increases in patients with IBS were also seen in the posterior insula (INS), the primary viscerosensory cortex, and these changes were correlated with IBS symptom duration.[91] Volumetric grey matter analyses in a large sample of female patients with IBS revealed increases in the primary somatosensory cortex.[78] Furthermore, examining structural networks in IBS indicated that two key regions of the network (cingulate gyrus and thalamus) were found to be network hubs, indicating that these regions are more critical for information flow in IBS compared with health.[78] When viewed together, current evidence supports the hypothesis that patients with chronically recurring visceral pain and/or discomfort have functional as well as neuroplastic and microstructural alterations within the brain, particularly in regions associated with the processing, integration and modulation of sensory information. The mechanism(s) underlying these alterations include chronically increased viscerosensory information flow from the gut, or from dorsal horn neurons sensitized by descending pain facilitation (see Figure 3).

Central executive network

Evidence indicates that patients with IBS might have functional impairments in cognitive processes associated with the central executive network.[16,17] Preliminary evidence based on administration of the Attentional Network Test[92] suggests that patients with IBS have greater behavioural efficiency during the alerting and orienting function of attention than healthy individuals. In comparison with healthy individuals, this greater efficiency was associated with greater activation of anterior midcingulate and insular cortices, confirming the previously reported close interactions between the central executive network and the salience network.[39] Converging evidence also suggests that increased attention to gastrointestinal symptoms and contexts have an important role in the increased perceptual sensitivity to visceral stimuli characteristic of IBS.[26] Patients with IBS show deficient activation of inhibitory cortical regions involved in downregulation of pain and emotion as well as attention during expectation and experience of aversive gastrointestinal stimuli.[53] Selective recall of negative and gastrointestinal sensation words, as well as selective attention to threat-related stimuli, has been demonstrated in patients with IBS.[93–96] Furthermore, a reduction in the effective connectivity of the central executive network circuitry (including parietal, dorsal lateral prefrontal cortex) during repeated exposure to the anticipation and experience of a threatening gastrointestinal stimulus (repeated exposure to balloon inflations) was associated with a reduction in IBS hypersensitivity.[97] Data from a sample of Japanese patients with IBS, compared with healthy controls, indicated that alterations in error feedback mechanisms were associated with decreased dorsolateral prefrontal cortex activity.[16] A strong negative correlation between the cortical thickness and grey matter density of the dorsolateral prefrontal cortex and pain catastrophizing has been reported.[98,99] Evidence from the University of California, Los Angeles research group indicates prepulse inhibition, a process by which an organism can filter the flow of information from its internal and external environments, is altered in IBS compared with health.[60] Together, these data suggest that patients with IBS have specific abnormalities in attentional processes that have a role in the increased perception of visceral stimuli and in IBS symptom severity.

Salience network

Studies performed during the past decade on brain responses to delivered and expected rectal distension have consistently reported increased engagement of the core regions of the salience network, the anterior INS and anterior midcingulate cortex in patients with IBS,[7,37] which initially but incorrectly were referred to as the "pain matrix". In addition, a close relationship between increased affect, central emotional arousal processes and enhanced visceral stimulus perception has been reported in patients with IBS.[65,100,101] Three recent reports published between 2013 and 2015 in female patients with IBS have identified disease-related alterations in anterior INS activity and connectivity in the resting state[58,59] and during an ambiguous abdominal pain threat,[102] confirming a key role of salience network alterations in IBS. Reported alterations in the response and connectivity within the salience network are consistent with the prediction error characteristic of patients with IBS about the likelihood and severity of future gastrointestinal symptoms (catastrophizing).[15]

Integration of brain networks

An extensive body of literature supports the model of the central role of aberrant salience computation underlying key clinical IBS symptoms (depicted in Figure 3). Well-established biological and behavioural consequences exist with such a biased appraisal: the engagement of altered ANS outputs to targets in the gastrointestinal tract (including ENS activity, gut permeability, gastrointestinal motility and secretion,[3] gut microbial composition and metabolites),[21,22] and to extraintestinal targets (including the immune system);[103] shifting of the balance of endogenous pain modulation systems towards increased descending pain facilitation;[20] increased engagement of the central executive network resulting in selective attention to gastrointestinal symptoms; and development of prediction errors about likelihood and severity of symptoms (so-called catastrophizing).[104,105] The model also provides a plausible biological basis for the effectiveness of different behavioural interventions such as cognitive behavioural therapy (normalizing salience, executive control and emotional arousal networks), self-relaxation techniques (normalizing emotional arousal and central autonomic networks) and mindfulness-based stress reduction (normalizing salience and executive control networks).[63]

© 2015 Macmillan Publishers Limited. All rights reserved

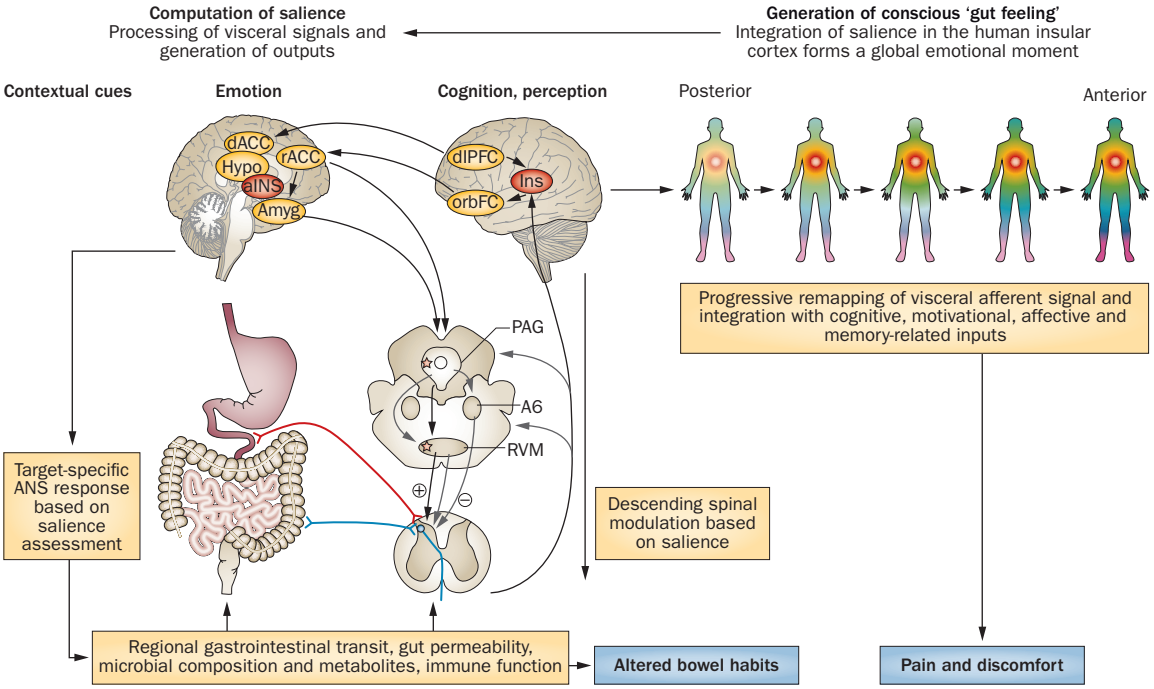


Figure 3 | Cross-sectional integrated brain–gut model of IBS pathophysiology. Proposed model for involvement of brain–gut axis in the generation of cardinal IBS symptoms (chronic abdominal pain associated with altered bowel habits). Under normal circumstances, visceral and external signals are evaluated by the salience network, which generates brain outputs in terms of targeted ANS responses (regulating gastrointestinal and immune function) and descending pain modulatory activity (regulating pain sensitivity at the dorsal horn level). Target organ alterations (either peripherally or ANS stimulated) are signalled back to the brain via neural, endocrine or immune-related channels. These signals are processed within subregions of the INS, and depending on their subjective salience, are consciously perceived (associated with activation of anterior INS) as normal gut sensations, discomfort or pain. IBS symptoms can arise from several primary peripheral or central mechanisms, but once brain–gut interactions are altered, causality is difficult to determine. Abbreviations: Amyg, amygdala; ANS, autonomic nervous system; dACC, dorsal anterior cingulate cortex; dlPFC, dorsolateral prefrontal cortex; Hypo, hypothalamus; INS, insula; orbFC, orbitofrontal cortex; PAG, periaqueductal grey; rACC, rostral anterior cingulate cortex; RVM, rostral ventromedial medulla.

Genetics and epigenetics

As with many other chronic diseases that involve the brain, IBS probably has a strong developmental component, which starts with the interactions of genetic and epigenetic factors early in life, including the prenatal period. In addition to the reported associations of gene variations with various peripheral mechanisms,[106] imaging genetics studies performed in large samples of well-phenotyped individuals with and without IBS have identified interactions between early environmental factors,[107] candidate gene polymorphisms and brain networks related to emotional arousal and/or central autonomic control, salience and somatosensory integration. The reported genes were related to the regulation of the HPA axis: corticotropin-releasing hormone receptor 1 (*CRHR1*, single nucleotide polymorphisms [SNPs] rs7209436, rs110402 and rs242924); glucocorticoid receptor gene (*NR3C1*, SNPs rs2963155 and rs33389);[108,109] female sex hormones (progesterone receptor or *PGR*, SNPs rs1042838 and rs10895068);[109] 5-HT signalling system (*HTR3A* c.–42C>T SNP rs1062613),[77] inflammation-related genes (*IL1B*, SNP rs16944);[108] and catecholaminergic signalling (*ADRA1D*, *SNP* rs1556832; *ADRA2B*, SNP rs1042717; *COMT*, SNP rs174697).[110] As discussed in the next section, these interactions are markedly influenced by epigenetic factors[111,112] (such as a history of early adverse life events [EALs]) and by the sex of the study participant (Figure 4).

As with other polygenic disorders, it has become clear that no single gene variation is sufficient to explain the full clinical phenotype in IBS. However, interactions between multiple genes, early life experiences and sex probably make a small contribution to the overall variance of the peripheral and central endophenotypes.[107] Validation studies in larger samples are required to confirm such contributions.

The external environment component

Various influences originating in the external environment (Figure 1) can markedly affect the development, chronicity and severity of IBS. Although some of these influences are mediated by the brain (psychosocial stress, social support, societal responses to symptoms) others are mediated by the gastrointestinal tract, including the gut microbiota, diet, gastrointestinal infections and medications.

Brain-mediated influences

Clinical literature on the role of psychosocial factors in the development, symptom persistence and symptom flares in IBS is extensive.[18] In addition to the well-documented

© 2015 Macmillan Publishers Limited. All rights reserved

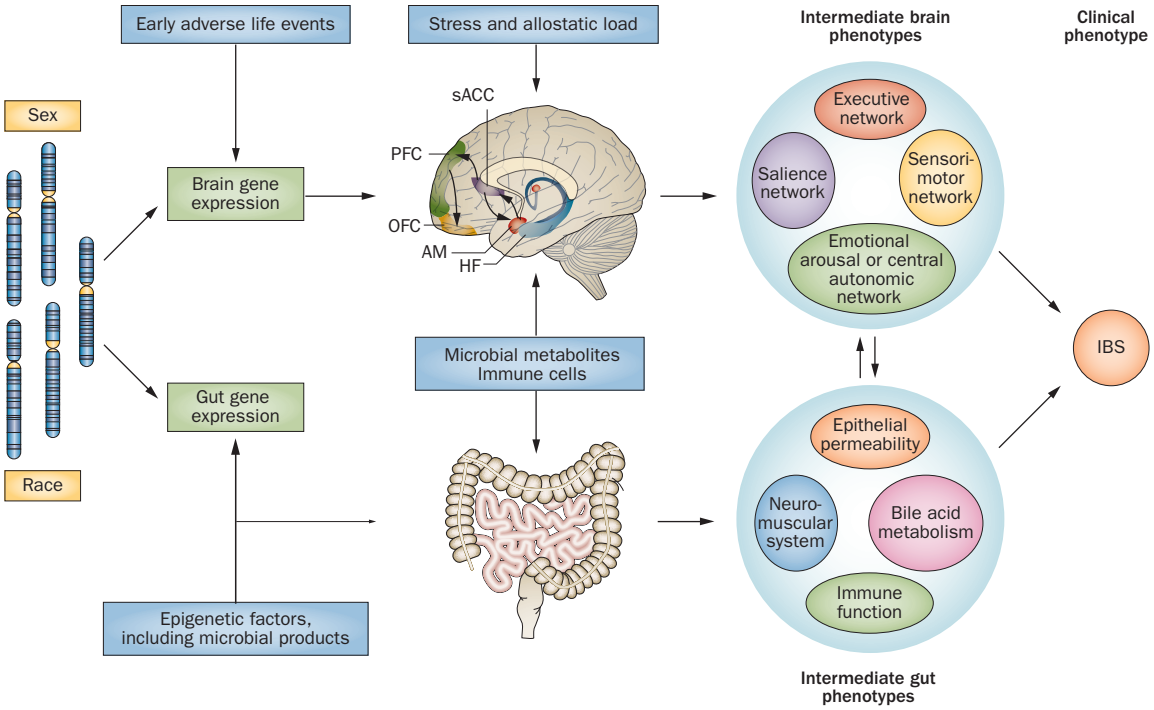


Figure 4 | Longitudinal brain–gut model of IBS pathophysiology. The interaction between genetic and epigenetic influences result in central and peripheral gene expression profiles that underlie the shaping of brain-based and gut-based intermediate phenotypes. Epigenetic factors provide the input from environmental influences on the development of intermediate phenotypes. Brain and gut intermediate phenotypes interact bidirectionally to shape the clinical phenotype of IBS. Abbreviations: AM, amygdala; HF, hippocampal formation; OFC, orbitofrontal cortex; PFC, prefrontal cortex; sACC, subgenual anterior cingulate cortex.

role of stressful life events in adults in preceding or exacerbating IBS symptoms,[32] a history of EALs is associated with an increased vulnerability to IBS,[113] disorders of mood and affect,[114,115] as well as to a wide range of other chronic diseases.[116] In contrast to the earlier emphasis on the role of a sexual abuse history, it is now clear that a variety of factors that disturb the quality of the interactions between the primary care giver and the child during the first 18 years of life (including serious illness of the mother, marital discord, divorce, verbal and emotional abuse) can have equally detrimental effect on adult disease vulnerability.[116]

Evidence from brain imaging studies have identified structural and functional brain alterations associated with self-reports of EALs. Structural alterations in regions of the emotional arousal circuitry[78] and in regions associated with modulating somatosensory and viscerosensory processes[90] were correlated with such reports. Alterations in the activity in the brainstem and amygdala in response to noradrenergic stimulation has been associated with increasing levels of EALs, consistent with an upregulation of central autonomic circuits in IBS.[117] Self-reports of EALs were also found to be correlated with brain activity in networks involved in determining the salience of somatic, visceral or environmental stimuli in IBS.[57] These results suggest that the experience of adversity early in life can lead to altered resting state activity in the salience and in the central executive network of adults with IBS, possibly leading to permanent alterations in salience computation of viscerosensory signals by the brain. The observed alterations in the functional connectivity of the emotional arousal[51] and salience networks[57] found in IBS might be driven by altered central noradrenergic modulation.[50,64] Similar changes have also been observed in animals exposed to early life stressors,[118] and have been linked to increased sympathetic nervous system responses.[119] EALs are also associated with altered signalling within the HPA axis.[120,121] A study published in 2009 demonstrated that self-report of EALs was associated with exaggerated HPA axis responses to an aversive visceral stimulus, an effect that was more pronounced in male participants.[113]

Considerable preclinical and clinical evidence supports the concept that gene expression can be influenced by EALs through epigenetic mechanisms, including DNA methylation, and that these effects can persist throughout adult life.[115,122,123] As depicted in Figure 4, interactions of EALs, sex and vulnerability gene polymorphisms might increase the risk of developing IBS by shaping the connectivity of relevant brain networks (see previous section). Even though the prevalence, clinical importance and underlying molecular mechanisms of EALs have been studied in great detail, there are other factors through which the external environment can influence brain function. These include, but are not limited to, the beneficial role of a strong social support system, which can mitigate the negative effects of EALs, and societal responses to patients' symptom reporting, which can drive a vicious cycle of symptom amplification.[124]

© 2015 Macmillan Publishers Limited. All rights reserved

Gastrointestinal-tract-mediated influences

Factors arising from the external environment have been implicated in the pathophysiology of IBS and in the modulation of IBS symptoms. These factors have been reviewed elsewhere and so will not be discussed in detail here: dietary factors;[125–128] pathogenic microorganisms;[129,130] and antibiotic treatment.[131,132] It remains to be determined if patients with IBS have abnormal mucosal responses to any of these factors, if symptoms reported in relation to these factors are mediated by alterations in the gut microbiota (see section on gut microbiota), or if it is simply the sensitivity of visceral perception that determines if somebody develops symptoms or can tolerate the same factors without any symptoms.

The immune system component

IBS is not an inflammatory disease, but a growing body of research suggests that dysregulation in immune function might nevertheless contribute to its aetiology or symptoms.[4,5,133] Mixed data exist on whether plasma or intestinal mucosal cytokine levels are associated with IBS.[5] However, other research has linked IBS to an increased reactivity of blood monocytes and increased numbers of mucosal mast cells.[134–139] Some studies also indicate that peripheral blood mononuclear cells (PBMCs) from patients with IBS show abnormal release of proinflammatory cytokines such as IL-6, IL-1b and TNF.[5]

These observations in IBS are consistent with the broader role of the nervous system in regulating immune cell development and gene expression via neuroendocrine signals from the brain (for example, cortisol from the HPA) and activation of sympathetic nerve fibres in the one marrow and other lymphoid organs (for example, via the sympathetic neuroeffector molecule norepinephrine).[103,140–142] These regulatory interactions enable the CNS to integrate information regarding the general internal state of the body with information regarding real and perceived environmental threats (as detected by the salience network) and historical or developmental influences (for example, a history of EALs[143]). As one example of these regulatory effects, studies have found that stressful life circumstances are associated with the activation of a so-called conserved transcriptional response to adversity (CTRA) in PBMCs that is characterized by increased expression of proinflammatory genes and decreased expression of genes involved in innate antiviral responses (for example, type I interferons) and IgG antibody production.[103,140,144,145] *CTRA* gene expression can also be experimentally invoked by social stress in animal models[145,146] and is mediated in part by sympathetic nervous system (SNS)-induced increases in bone marrow haematopoietic production of immature and immunologically primed monocytes ($CD16^-$ in humans, Ly-$6c^{hi}$ in mice).[146] These primed monocytes can also be reciprocally recruited into the CNS by exposure of mice to social threat and aggression.[147] The integration of these observations suggests a new hypothesis regarding the immune system's role in the pathogenesis of IBS: high levels of SNS activity during early developmental periods (stemming from either genetic or environmental triggers such as EALs) might lead to increased production of immature primed monocytes that both traffic into the gut to alter local function and ENS plasticity, and traffic into the brain to affect CNS plasticity (including structures involved in salience processing and autonomic regulation). The result can be viewed as a physiological Hebbian association between gut biology and brain function, resulting in a self-perpetuating feedback system in which a sensitized gut generates ongoing adverse sensory experiences (symptom flares), to which a neurally sensitized brain responds with both greater aversion and increased sympathetic outflow, resulting in upregulated monocyte production that further promotes neural alterations in both the gut and the brain.

Several areas of empirical evidence are consistent with this systems-level brain–immune–gut hypothesis, including: brain regulation of primed monocyte production via the SNS;[148] stress-induced migration of primed monocytes to the brain;[147] inflammation-induced neuroplastic changes in the brain[149] (for example, altering affective behaviour[147]); reinstatement of previous stress effect on brain and behaviour by subsequent exposure to mild stress weeks after the initial sensitization;[148] and the observation that patients with IBS (and animal models of EAL) show increased stress responsiveness (for example, SNS activity[150]) and increased responsiveness of brain circuits related to salience detection, emotional arousal and autonomic response.[151,152]

Additional evidence was examined from a pilot study in which the PBMC gene expression profiles from 20 patients with IBS (12 female) and 20 healthy individuals (nine female).[153] Analyses identified 280 gene transcripts showing >10% differential expression across groups (134 genes upregulated in PBMC from IBS, and 146 downregulated). Promoter-based bioinformatics analysis implicated several transcription control pathways in structuring the observed transcriptome differences, including increased activity of CREB transcription factors (which mediate β-adrenergic signalling from the SNS), growth control pathways (for example, the MAPK-responsive transcription factor ELK1), oxidative stress response pathways (NRF2), and pathways involved in growth factor and cytokine signalling (STAT). However, PBMCs derived from patients with IBS did not differ from those from healthy individuals in the activity of proinflammatory transcription factors such as NFκB or AP-1. Transcript origin analyses[154] indicated that IBS upregulated genes derived predominately from monocytes and dendritic cells. Additional transcriptome representation analyses suggest that these effects stemmed at least in part from upregulation of immature ($CD16^-$) monocytes within the PBMC pool of patients with IBS. These results are all consistent with a pattern of increased myeloid lineage cell development in patients with IBS, which might stem from tonically increased SNS signalling to bone marrow myelopoietic processes.[146]

Consistent with the hypothesis that CNS processes might mediate relationships between IBS and peripheral myelopoiesis, differential signalling by the myeloid lineage transcription factor MZF-1 was positively

© 2015 Macmillan Publishers Limited. All rights reserved

associated with the morphometry of brain regions of the emotional arousal network[155] (previously reported to be altered in IBS[78,90,156]). These associations included bilateral amygdala, hippocampal and anterior INS volumes, and bilateral anterior INS cortical thickness (all $P<0.05$). Moreover, these associations were reasonably specific in that other brain regions showed either no association, or negative associations (for example, bilateral cerebellar volumes).

In conjunction with the published data reviewed here, these preliminary results suggest that key regions of the emotional arousal and/or central autonomic networks, which differ structurally between IBS and healthy individuals are also related to differential gene regulation in the peripheral immune system. We hypothesize that these chronic influences of the brain on peripheral monocytes ('top down') might have a role in the reported PBMC abnormalities in IBS, whereas primed monocytes migrating to the brain might have a role in the ('bottom up') generation of visceral hypersensitivity, anxiety and neuroplasticity during recurrent stressors.

The gut microbiota component

Evidence from rodent studies supporting bidirectional interactions between the gut microbiota and the nervous system (both CNS and ENS) has been summarized in numerous review articles.[157–160] Even though various signalling mechanisms underlying such interactions have been proposed, detailed mechanistic studies regarding the relative contributions of neural, hormonal, metabolite or immune-mediated factors are required to draw definitive conclusions. Evidence from human studies using different endpoints (symptoms, brain imaging studies) confirming the rodent findings, or identifying a definitive pattern of dysbiosis in patients with IBS are limited.[158]

Conflicting evidence exists regarding alterations to the organization and function of the gut microbiota in patients with chronic abdominal pain and in adult and paediatric patients with IBS.[21,22] Several studies examining faecal samples from patients with IBS reported decreased proportions of the genera *Bifidobacterium* and *Lactobacillus*, and increased ratios of Firmicutes:Bacteroidetes at the phylum level, even though a causal role of these microbial changes in clinical symptoms has not been established. On the other hand, one might speculate that some of these changes might be related to alterations in regional gut transit and secretion secondary to altered ANS output. A study in a cohort of well-phenotyped patients with IBS highlights the complexities of brain–gut–microbiota alterations in IBS.[31] IBS subgroups were identified based on hierarchical clustering of operational taxonomic unit information from 16S ribosomal RNA analyses.[31] Two IBS clusters were clearly separated from each other, from a 'normal-like' IBS sample and from the healthy control sample. Although the normal-like IBS sample showed normal diversity and normal Firmicutes:Bacteroidetes ratio (as in, similar to the pattern in the healthy control sample), another cluster showed diminished, and a third cluster showed increased diversity. Both of these clusters showed increased Firmicutes:Bacteroidetes ratios. Only the normal-like IBS group showed a markedly elevated rate of depression symptoms (40%). In addition, prolonged colonic transit times, common in patients with depression correlated with the prevalence of 17 taxa.[31] In a randomized, placebo-controlled study of healthy men and women, psychological distress and anxiety improved after taking a *Lactobacillus*-containing and *Bifidobacterium*-containing probiotic compared to those taking a matched control product, although another study using a different *Lactobacillius* probiotic failed to confirm these findings.[161,162]

Findings from brain imaging studies in healthy individuals and patients with IBS has provided some evidence supporting reported rodent gut microbiota–brain interactions. One study has shown that chronic ingestion of a probiotic consortium (*Bifididobacterium animalis* ssp. *lactis*, *Lactococcus lactis* spp. *lactis*, *Lactobacillus delbrueckii* spp. *bulgaricus* and *Streptococcus thermophilus*) for 4 weeks altered functional brain responses to an emotional face recognition task in healthy women.[52] Compared with two control groups (one group received a nonfermented milk product and the other no treatment), the women who had ingested the probiotic had a reduced response to the task across a wide network of brain regions that included sensory and emotional regions. Although no treatment-related changes in self-report of symptoms of anxiety or depression were seen, the findings suggest a basic change in responsiveness to negative emotional stimuli in the environment. No organizational changes in the gut microbiota were observed in this and a previous nonimaging study using the same intervention.[163] The effects of the probiotic intervention on brain responsiveness were probably mediated by a change in microbial-derived metabolites; this hypothesis will have to be confirmed in future studies. Preliminary evidence investigating correlations between gut microbial metabolites and brain structure in healthy individuals and patients with IBS demonstrated statistically significant correlations between several metabolites and structural aspects of several brain regions.[153]

An integrated model

Hypothesis

On the evidence reviewed here, the following model is proposed that lends itself to experimental evaluation. The hypersensitive brain, and presumably ENS, shows increased responses to a variety of viscerosensory and exterosensory stimuli, which by themselves might not be consciously perceived in healthy individuals or in patients with inflammatory bowel disorders in whom intact descending inhibitory bulbospinal influences reduce dorsal horn excitability.[20,164] This hyper-responsiveness might be a primary genetic or epigenetic alteration in certain brain networks (autonomic, emotional arousal or salience), or might be secondary to chronic experience of increased sensory input from the gut. Such increased viscerosensory signalling could originate from any of the elements of the gut connectome. Altered brain networks generate altered signals, which are transmitted to the

© 2015 Macmillan Publishers Limited. All rights reserved

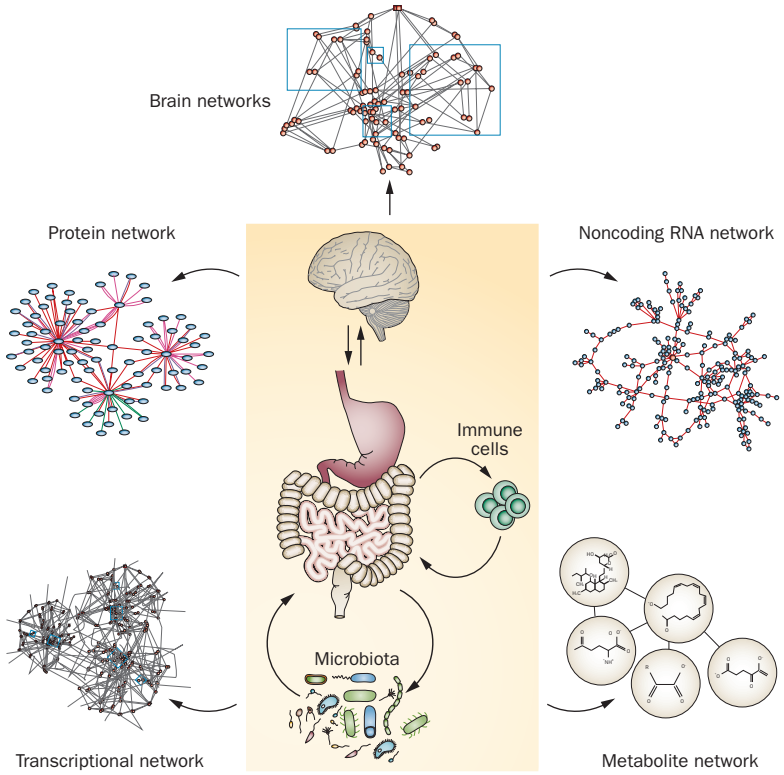


Figure 5 | Systems biological model of IBS. Schematic illustrating a systems biological view of components involved in the development of IBS at the cellular and molecular level. Systems-based interactions between central and peripheral components of the brain–gut axis can be studied at the level of the genome, epigenome, transcriptome, proteome, metabolome and brain connectome.

periphery through the ANS, HPA axis and descending modulation to the dorsal horn. The chronically increased ANS output, results in an extensive remodelling of various peripheral cells in the immune system, ENS, gut epithelium (permeability), and in the composition and function of the gut microbiota, all contributing to sensitization of visceral afferent pathways, and increased viscerosensory feedback to the brain, reinforcing the circular regulatory loops. Memory formations at the level of the immune system, the nervous system and the gut microbiota early in life are likely to contribute to the chronicity of symptoms.

Systems biology view

The evidence reviewed above strongly suggests that IBS is a systems disease (Figure 5), involving not only complex individual systems—nervous, immune, digestive, microbiota and the environment—but also their complex, nonlinear, reciprocal interactions. Thus, we propose a systems biological view of IBS pathophysiology, an approach taken in other areas of biology and disease[165,166] involving a large number of functionally diverse, interacting components, each of which contributes only a small fraction to the variance of the symptoms. These components interact highly selectively at multiple scales and typically in a nonlinear fashion to produce coherent behaviours and outcomes.[167] Applied on a macroscopic level, we propose this model to characterize the interactions of the nervous system, immune system, the gut environment (motility and secretion), the gut microbiota and the external environment (Figures 1 and 5). The nonlinearity of the model is reflected by the complexity of these macroscopic components and their macroscopic interactions, which form circular regulatory loops. Applied to the microscopic level, the model proposes the interactions between highly diverse cells types making up the various macroscopic components (Figure 5).

Given the complexity of the expanded gut–microbiota–brain–environment axis, and with the rapid advance of analytical tools, in particular of '-omics' technologies—in the case of IBS studies, primarily genomics, transcriptomics, proteomics, metabolomics, brain connectomics and their integration into, for instance, gene and protein networks—we believe that systems biological approaches are essential to fully understand symptom generation and to identify novel treatment approaches in the future.

The multiorgan, systemic view of IBS leads to the prediction that disease models focusing on individual cellular components at the macroscopic level (brain, immune system, gut, microbiota, environment) will have limited validity. Furthermore, this view predicts that therapies targeted at single organs (for example, brain, immune system, gut), single mechanisms (for example, secretion, motility, gut microbiota, pain, diet), or single molecular targets (such as ion channels or receptors) are not likely to be successful in treating the entire syndrome, as evidenced by the limited success in drug development for IBS to date, reflected in rather small effect sizes on overall changes in IBS symptoms. This problem is compounded by the fact that different subsets of patients can be characterized by different patterns of interactions between the components of the overall system, resulting in differential responses to a particular therapy.

Indeed, each isolated mechanism proposed over the past 30 years only explains a small fraction of the variance of the clinical phenotype[26] and no isolated therapies have had better effectiveness than ~10% over placebo.[168] In addition, the systems view predicts that a single agent targeting multiple components of the system (tricyclic antidepressants, 5-HT_3 antagonists), or combined therapies, aimed at simultaneously targeting multiple organs and mechanisms (combination of laxatives, antidiarrhoeal agents, probiotics and centrally acting drugs) are likely to be more effective in the clinic, than individual treatments by themselves. Although no controlled studies have been conducted yet and are needed, in our opinion, this suggested integrated approach is in part supported by current clinical practices and the consensus of most physicians.

Although still in their infancy, integrated multiomic approaches could be essential in the future to better understand the disease spectrum for IBS at the system level and the underlying mechanisms. Ideally, one would like to combine high-throughput microscopy and brain imaging to obtain brain connectome variables, and sequencing and mass spectrometry methods to obtain complete measurements of genomic, epigenomic, transcriptomic, proteomic and metabolomic variables simultaneously in multiple organs and tissues (in particular in gut and brain), including the microbiota. Data derived from these

© 2015 Macmillan Publishers Limited. All rights reserved

–omics approaches can be integratively analysed in combination with other sources of information ranging from public databases to the literature to identify, for instance, targetable hubs. This integrated –omics approach faces major obstacles stemming from the obvious difficulties of performing some of these analyses in human patients (for example, nervous system tissue). However, important steps can be taken in this direction both using rodent models[10,81,169,170] and by performing partial measurements in humans.

Some of these analyses should be performed at multiple time points along the circadian cycle. Circadian rhythms are found at the molecular level in all tissues, both centrally and peripherally, and have a fundamental role in coordinating the physiology and homeostasis of the organism. Indeed integrative system biology analyses[171] have revealed that about 10% of all transcripts and metabolites oscillate in a circadian manner in any tissue,[172] with little overlap across different tissues beyond the core molecular clock comprising a transcriptional–translational negative feedback loop coordinated by a dozen genes.[173] Furthermore, the list of oscillating molecular species in a given tissue is altered by genetic, epigenetic and environmental perturbations and thus provides a characteristic physiological signature of a tissue and its health state.[174–176] Furthermore, complex reciprocal interactions exist between the molecular rhythms found in different brain regions and in other organs. Thus, in short, it is reasonable to predict that IBS should result in perturbed lists of transcripts and metabolites that oscillate in a circadian manner both centrally in specific brain areas as well as peripherally, for instance in the gut.

Conclusions

Clearly, none of the growing list of individual abnormalities identified in patients with IBS by itself can account for the variance of the clinical phenotype. For the same reason, such individual abnormalities are unlikely to represent reliable biomarkers and are not likely to represent suitable targets for the development of highly effective treatments. Rather, as depicted schematically in Figures 1 and 5, the clinical phenotype emerges from the interactions of multiple systems in the periphery (gut connectome, microbiome, genome and epigenome) and in the brain (connectome) interacting with each other in bidirectional ways. Most consistent with a systems view is the concept that central and peripheral abnormalities form circular loops that reinforce each other. In the absence of comprehensive phenotyping studies in patients with IBS performed longitudinally with or without therapeutic interventions, and without targeted mechanistic animal studies, it remains unclear which of these reported abnormalities are primary and which are secondary. On the basis of this assessment, we suggest that in the future, high-throughput –omics measurement across both tissues and time, combined with comprehensive characterization of clinical, behavioural and brain endophenotypes, should enable more accurate differential analyses, uncover complex system-level interactions and, ultimately, help develop more efficient, multipronged therapies.[177] Such a system biological approach might not only hold promise when applied to IBS and related functional gastrointestinal disorders, but also for inflammatory diseases of the gut, such as IBD and coeliac disease.[178]

1. Longstreth, G. F. *et al.* in *ROME III: The Functional Gastrointestinal Disorders* (eds Drossman, D. A. *et al.*) 487–556 (Degnon Associates, 2006).
2. Mayer, E. A. & Bushnell, M. C. in *Functional Pain Syndromes: Presentation and Pathophysiology* (eds Mayer, E. A. & Bushnell, M. C.) 531–565 (IASP Press, 2009).
3. Camilleri, M. Peripheral mechanisms in irritable bowel syndrome. *N. Engl. J. Med.* **367**, 1626–1635 (2012).
4. Hughes, P. A. *et al. Immune* activation in irritable bowel syndrome: can neuroimmune interactions explain symptoms? *Am. J. Gastroenterol.* **108**, 1066–1074 (2013).
5. Ohman, L. & Simren, M. Pathogenesis of IBS: role of inflammation, immunity and neuroimmune interactions. *Nat. Rev. Gastroenterol. Hepatol.* **7**, 163–173 (2010).
6. Fukudo, S. Stress and visceral pain: focusing on irritable bowel syndrome. *Pain* **154** (Suppl. 1), S63–S70 (2013).
7. Mayer, E. A., Gupta, A., Kilpatrick, L. K. & Hong, J. Y. Chronic visceral pain. *Pain* **156** (Suppl. 1), S50–S63 (2015).
8. Elsenbruch, S. Abdominal pain in irritable bowel syndrome: a review of putative psychological, neural and neuro-immune mechanisms. *Brain Behav. Immun.* **25**, 386–394 (2011).
9. Al Omran, Y. & Aziz, Q. Functional brain imaging in gastroenterology: to new beginnings. *Nat. Rev. Gastroenterol. Hepatol.* **11**, 565–576 (2014).
10. Greenwood-Van Meerveld, B., Prusator, D. K. & Johnson, A. C. Animal models of visceral pain: pathophysiology, translational relevance and challenges. *Am. J. Physiol. Gastrointest. Liver Physiol.* **308** , G885–G903 (2015).
11. Lackner, J. M. in *Functional and GI Motility Disorders* (eds Quigley, E. M. M., Hongo, M. & Fukudo, S.) 104–116 (Karger Medical and Scientific Publishers, 2014).
12. Bischoff, S. C. *et al.* Intestinal permeability —a new target for disease prevention and therapy. *BMC Gastroenterol.* **14**, 189 (2014).
13. Mawe, G. M. & Hoffman, J. M. Serotonin signalling in the gut—functions, dysfunctions and therapeutic targets. *Nat. Rev. Gastroenterol. Hepatol.* **10**, 473–486 (2013).
14. Larauche, M., Mulak, A. & Tache, Y. Stress and visceral pain: from animal models to clinical therapies. *Exp. Neurol.* **233**, 49–67 (2012).
15. Elsenbruch, S. How positive and negative expectations shape the experience of visceral pain. *Handb. Exp. Pharmacol.* **225**, 97–119 (2014).
16. Aizawa, E. *et al.* Altered cognitive function of prefrontal cortex during error feedback in patients with irritable bowel syndrome, based on FMRI and dynamic causal modeling. *Gastroenterology* **143**, 1188–1198 (2012).
17. Kennedy, P. J. *et al.* Cognitive performance in irritable bowel syndrome: evidence of a stress-related impairment in visuospatial memory. *Psychol. Med.* **44**, 1553–1566 (2014).
18. Tanaka, Y., Kanazawa, M., Fukudo, S. & Drossman, D. A. Biopsychosocial model of irritable bowel syndrome. *J. Neurogastroenterol. Motil.* **17**, 131–139 (2011).
19. Piche, M., Arsenault, M., Poitras, P., Rainville, P. & Bouin, M. Widespread hypersensitivity is related to altered pain inhibition processes in irritable bowel syndrome. *Pain* **148**, 49–58 (2010).
20. Wilder-Smith, C. H. The balancing act: endogenous modulation of pain in functional gastrointestinal disorders. *Gut* **60**, 1589–1599 (2011).
21. Mayer, E. A., Savidge, T. & Shulman, R. J. Brain–gut microbiome interactions and functional bowel disorders. *Gastroenterology* **146**, 1500–1512 (2014).
22. Simren, M. *et al.* Intestinal microbiota in functional bowel disorders: a Rome foundation report. *Gut* **62**, 159–176 (2013).
23. Drossman, D. A. in *The* Growth *of Gastroenterologic Knowledge in the 20th Century* (ed. Kirsner, J. B.) 419–432 (Lea & Febiger, 1993).
24. Mayer, E. A., Tillisch, K. & Bradesi, S. Review article: modulation of the brain-gut axis as a therapeutic approach in gastrointestinal disease. *Aliment. Pharmacol. Ther.* **24**, 919–933 (2006).
25. Barboza, J. L., Talley, N. J. & Moshiree, B. Current and emerging pharmacotherapeutic options for irritable bowel syndrome. Drugs **74**, 1849–1870 (2014).
26. Mayer, E. A. *et al.* Functional GI disorders: from animal models to drug development. *Gut* **57**, 384–404 (2008).
27. Keita, A. V. & Soderholm, J. D. The intestinal barrier and its regulation by neuroimmune factors. *Neurogastroenterol. Motil.* **22**, 718–733 (2010).

© 2015 Macmillan Publishers Limited. All rights reserved

28. Elenkov, I. J. & Chrousos, G. P. Stress system—organization, physiology and immunoregulation. *Neuroimmunomodulation* **13**, 257–267 (2006).
29. Hayes, P. A., Fraher, M. H. & Quigley, E. M. Irritable bowel syndrome: the role of food in pathogenesis and management. *Gastroenterol. Hepatol. (N. Y.)* **10**, 164–174 (2014).
30. Spiller, R. & Lam, C. An Update on post-infectious irritable bowel syndrome: role of genetics, *immune* activation, serotonin and altered microbiome. *J. Neurogastroenterol. Motil.* **18**, 258–268 (2012).
31. Jeffery, I. B. *et al.* An irritable bowel syndrome subtype defined by species-specific alterations in faecal microbiota. *Gut* **61**, 997–1006 (2012).
32. Mayer, E. A. The neurobiology of stress and gastrointestinal disease. *Gut* **47**, 861–869 (2000).
33. Ford, A. C. *et al.* Effect of antidepressants and psychological therapies, including hypnotherapy, in irritable bowel syndrome: systematic review and meta-analysis. *Am. J. Gastroenterol.* **109**, 1350–1365 (2014).
34. Mayer, E. A. Gut feelings: the emerging biology of gut–brain communication. *Nat. Rev. Neurosci.* **12**, 453–466 (2011).
35. Furness, J. B. The enteric nervous system and neurogastroenterology. *Nat. Rev. Gastroenterol. Hepatol.* **9**, 286–294 (2012).
36. Bohorquez, D. V. & Liddle, R. A. The gut connectome: making sense of what you eat. *J. Clin. Invest.* **125**, 888–890 (2015).
37. Mayer, E. A. *et al.* Brain imaging approaches to the study of functional GI disorders: a Rome working team report. *Neurogastroenterol. Motil.* **21**, 579–596 (2009).
38. Farmer, A. D., Aziz, Q., Tack, J. & Van Oudenhove, L. The future of neuroscientific research in functional gastrointestinal disorders: integration towards multidimensional (visceral) pain endophenotypes? *J. Psychosom. Res.* **68**, 475–481 (2010).
39. Seeley, W. W. *et al.* Dissociable intrinsic connectivity networks for salience processing and executive control. *J. Neurosci.* **27**, 2349–2356 (2007).
40. Guo, C. C. *et al.* One-year test-retest reliability of intrinsic connectivity network fMRI in older adults. *Neuroimage* **61**, 1471–1483 (2012).
41. Bajaj, S., Adhikari, B. M. & Dhamala, M. Higher frequency network activity flow predicts lower frequency node activity in intrinsic low-frequency BOLD fluctuations. *PLoS ONE* **8**, e64466 (2013).
42. Bullmore, E. & Sporns, O. The economy of brain network organization. *Nat. Rev. Neurosci.* **13**, 336–349 (2012).
43. Rubinov, M. & Sporns, O. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage* **52**, 1059–1069 (2010).
44. Hutchison, R. M. *et al.* Resting-state connectivity identifies distinct functional networks in macaque cingulate cortex. *Cereb. Cortex* **22**, 1294–1308 (2012).
45. Greicius, M. D., Supekar, K., Menon, V. & Dougherty, R. F. Resting-state functional connectivity reflects structural connectivity in the default mode network. *Cereb. Cortex* **19**, 72–78 (2009).
46. Sporns, O. *Networks of the Brain* (The MIT Press, 2010).
47. Tijms, B. M., Series, P., Willshaw, D. J. & Lawrie, S. M. Similarity-based extraction of individual networks from gray matter MRI scans. *Cereb. Cortex* **22**, 1530–1541 (2012).
48. Hagmann, P. *et al.* Mapping human whole-brain structural networks with diffusion MRI. *PLoS ONE* **2**, e597 (2007).
49. Bradford, K. *et al.* Association between early adverse life events and irritable bowel syndrome. *Clin. Gastroenterol. Hepatol.* **10**, 385–390.e1–e3 (2012).
50. Dickhaus, B. *et al.* Irritable bowel syndrome patients show enhanced modulation of visceral perception by auditory stress. *Am. J. Gastroenterol.* **98**, 135–143 (2003).
51. Labus, J. S. *et al.* Sex differences in emotion-related cognitive processes in irritable bowel syndrome and healthy control subjects. *Pain* **154**, 2088–2099 (2013).
52. Tillisch, K. *et al.* Consumption of fermented milk product with probiotic modulates brain activity. *Gastroenterology* **144**, 1394–1401.e1–4 (2013).
53. Tillisch, K., Mayer, E. A. & Labus, J. S. Quantitative meta-analysis identifies brain regions activated during rectal distension in irritable bowel syndrome. *Gastroenterology* **140**, 91–100 (2011).
54. Farmer, A. D. & Aziz, Q. Visceral pain hypersensitivity in functional gastrointestinal disorders. *Br. Med. Bull.* **91**, 123–136 (2009).
55. Mayer, E. A., Berman, S., Chang, L. & Naliboff, B. D. Sex-based differences in gastrointestinal pain. *Eur. J. Pain* **8**, 451–463 (2004).
56. Tillisch, K. *et al.* Sex specific alterations in autonomic function among patients with irritable bowel syndrome. *Gut* **54**, 1396–1401 (2005).
57. Gupta, A. *et al.* Early adverse life events and resting state neural networks in patients with chronic abdominal pain: evidence for sex differences. *Psychosom. Med.* **76**, 404–412 (2014).
58. Hong, J. Y. *et al.* Sex and disease-related alterations of anterior insula functional connectivity in chronic abdominal pain. *J. Neurosci.* **34**, 14252–14259 (2014).
59. Hong, J. Y. *et al.* Patients with chronic visceral pain show sex-related alterations in intrinsic oscillations of the resting brain. *J. Neurosci.* **33**, 11994–12002 (2013).
60. Kilpatrick, L. A. *et al.* Sex-related differences in prepulse inhibition of startle in irritable bowel syndrome (IBS). *Biol. Psychol.* **84**, 272–278 (2010).
61. Van Oudenhove, L., Vandenberghe, J., Vos, R., Holvoet, L. & Tack, J. Factors associated with co-morbid irritable bowel syndrome and chronic fatigue-like symptoms in functional dyspepsia. *Neurogastroenterol. Motil.* **23**, 524–e202 (2011).
62. Naliboff, B. D. *et al.* Longitudinal change in perceptual and brain activation response to visceral stimuli in irritable bowel syndrome patients. *Gastroenterology* **131**, 352–365 (2006).
63. Naliboff, B. N., Lackner, J. & Mayer, E. A. in *Principles of Clinical Gastroenterology* (eds Yamada, T. *et al.*) (Wiley-Blackwell Publishing 2008).
64. Hubbard, C. S. *et al.* Corticotropin-releasing factor receptor 1 antagonist alters regional activation and effective connectivity in an emotional-arousal circuit during expectation of abdominal pain. *J. Neurosci.* **31**, 12491–12500 (2011).
65. Hall, G. B. *et al.* Heightened central affective response to visceral sensations of pain and discomfort in IBS. *Neurogastroenterol. Motil.* **22**, 276–e80 (2010).
66. Labus, J. S. *et al.* Acute tryptophan depletion alters the effective connectivity of emotional arousal circuitry during visceral stimuli in healthy women. *Gut* **60**, 1196–1203 (2011).
67. Straube, T., Schmidt, S., Weiss, T., Mentzel, H. J. & Miltner, W. H. Dynamic activation of the anterior cingulate cortex during anticipatory anxiety. *Neuroimage* **44**, 975–981 (2009).
68. Porro, C. A. *et al.* Does anticipation of pain affect cortical nociceptive systems? *J. Neurosci.* **22**, 3206–3214 (2002).
69. Koyama, T., McHaffie, J. G., Laurienti, P. J. & Coghill, R. C. The subjective experience of pain: where expectations become reality. *Proc. Natl Acad. Sci. USA* **102**, 12950–12955 (2005).
70. Bornhovd, K. *et al.* Painful stimuli evoke different stimulus-response functions in the amygdala, prefrontal, insula and somatosensory cortex: a single-trial fMRI study. *Brain* **125**, 1326–1336 (2002).
71. Verne, G. N. *et al.* Central representation of visceral and cutaneous hypersensitivity in the irritable bowel syndrome. *Pain* **103**, 99–110 (2003).
72. Seifert, F. *et al.* Brain activity during sympathetic response in anticipation and experience of pain. *Hum. Brain Mapp.* **34**, 1768–1782 (2013).
73. Yaguez, L. *et al.* Brain response to visceral aversive conditioning: a functional magnetic resonance imaging study. *Gastroenterology* **128**, 1819–1829 (2005).
74. Berman, S. M. *et al.* Reduced brainstem inhibition during anticipated pelvic visceral pain correlates with enhanced brain response to the visceral stimulus in women with irritable bowel syndrome. *J. Neurosci.* **28**, 349–359 (2008).
75. Stephan, E. *et al.* Functional neuroimaging of gastric distention. *J. Gastrointest Surg.* **7**, 740–749 (2003).
76. Lawal, A. *et al.* Neurocognitive processing of esophageal central sensitization in the insula and cingulate gyrus. *Am. J. Physiol. Gastrointest. Liver Physiol.* **294**, G787–G794 (2008).
77. Kilpatrick, L. A. *et al.* The HTR3A polymorphism c. –42C>T is associated with amygdala responsiveness in patients with irritable bowel syndrome. *Gastroenterology* **140**, 1943–1951 (2011).
78. Labus, J. S. *et al.* Irritable bowel syndrome in female patients is associated with alterations in structural brain networks. *Pain* **155**, 137–149 (2014).
79. de Kloet, E. R., Joels, M. & Holsboer, F. Stress and the brain: from adaptation to disease. *Nat. Rev. Neurosci.* **6**, 463–475 (2005).
80. McEwen, B. S. & Morrison, J. H. The brain on stress: vulnerability and plasticity of the prefrontal cortex over the life course. *Neuron* **79**, 16–29 (2013).
81. Holschneider, D. P., Bradesi, S. & Mayer, E. A. The role of experimental models in developing new treatments for irritable bowel syndrome. *Expert Rev. Gastroenterol. Hepatol.* **5**, 43–57 (2011).
82. Tache, Y., Martinez, V., Wang, L. & Million, M. CRF1 receptor signaling pathways are involved in stress-related alterations of colonic function and viscerosensitivity: implications for irritable bowel syndrome. *Br. J. Pharmacol.* **141**, 1321–1330 (2004).
83. Labus, J. S. *et al.* Impaired emotional learning and involvement of the corticotropin-releasing factor signaling system in patients with irritable bowel syndrome. *Gastroenterology* **145**, 1253–1261 e1–3 (2013).
84. Rubio, A. *et al.* Uncertainty in anticipation of uncomfortable rectal distension is modulated by the autonomic nervous system—a fMRI study in healthy volunteers. *Neuroimage* **107C**, 110–22 (2014).
85. Kilpatrick, L. A. *et al.* Alterations in resting state oscillations and connectivity in sensory and motor networks in women with interstitial cystitis/painful bladder syndrome. *J. Urol.* **192**, 947–955 (2014).
86. Apkarian, A. V., Thomas, P. S., Krauss, B. R. & Szeverenyi, N. M. Prefrontal cortical hyperactivity in patients with sympathetically mediated chronic pain. *Neurosci. Lett.* **311**, 193–197 (2001).

© 2015 Macmillan Publishers Limited. All rights reserved

87. Becerra, L. *et al.* Intrinsic brain networks normalize with treatment in pediatric complex regional pain syndrome. *Neuroimage Clin.* **6**, 347–369 (2014).
88. DaSilva, A. F. *et al.* Colocalized structural and functional changes in the cortex of patients with trigeminal neuropathic pain. *PLoS ONE* **3**, e3396 (2008).
89. Ellingson, B. M. *et al.* Diffusion tensor imaging detects microstructural reorganization in the brain associated with chronic irritable bowel syndrome. *Pain* **154**, 1528–1541 (2013).
90. Jiang, Z. *et al.* Sex-related differences of cortical thickness in patients with chronic abdominal pain. *PLoS ONE* **8**, e73932 (2013).
91. Piche, M. *et al.* Thicker posterior insula is associated with disease duration in women with irritable bowel syndrome (IBS) whereas thicker orbitofrontal cortex predicts reduced pain inhibition in both IBS patients and controls. *J. Pain* **14**, 1217–1226 (2013).
92. Fan, J., McCandliss, B. D., Fossella, J., Flombaum, J. I. & Posner, M. I. The activation of attentional networks. *Neuroimage* **26**, 471–479 (2005).
93. Afzal, M., Potokar, J. P., Probert, C. S. & Munafo, M. R. Selective processing of gastrointestinal symptom-related stimuli in irritable bowel syndrome. *Psychosom. Med.* **68**, 758–761 (2006).
94. Gibbs-Gallagher, N. *et al.* Selective recall of gastrointestinal-sensation words: evidence for a cognitive-behavioral contribution to irritable bowel syndrome. *Am. J. Gastroenterol.* **96**, 1133–1138 (2001).
95. Phillips, K., Wright, B. J. & Kent, S. Irritable bowel syndrome and symptom severity: Evidence of negative attention bias, diminished vigour, and autonomic dysregulation. *J. Psychosom Res.* **77**, 13–19 (2014).
96. Tkalcic, M., Domijan, D., Pletikosic, S., Setic, M. & Hauser, G. Attentional biases in irritable bowel syndrome patients. *Clin. Res. Hepatol. Gastroenterol.* **38**, 621–628 (2014).
97. Labus, J. S. *et al.* Brain networks underlying perceptual habituation to repeated aversive visceral stimuli in patients with irritable bowel syndrome. *Neuroimage* **47**, 952–960 (2009).
98. Blankstein, U., Chen, J., Diamant, N. E. & Davis, K. D. Altered brain structure in irritable bowel syndrome: potential contributions of pre-existing and disease-driven factors. *Gastroenterology* **138**, 1783–1789 (2010).
99. Seminowicz, D. A. *et al.* Cognitive-behavioral therapy increases prefrontal cortex gray matter in patients with chronic pain. *J. Pain* **14**, 1573–1584 (2013).
100. Elsenbruch, S. *et al.* Patients with irritable bowel syndrome have altered emotional modulation of neural responses to visceral stimuli. *Gastroenterology* **139**, 1310–1319 (2010).
101. Elsenbruch, S. *et al.* Affective disturbances modulate the neural processing of visceral pain stimuli in irritable bowel syndrome: an fMRI study. *Gut* **59**, 489–495 (2010).
102. Jui-Yang Hong *et al.* IBS patients show altered brain responses during uncertain, but not certain expectation of painful stimulation of the abdominal wall [abstract]. *Gastroenterology* **148** (Suppl.) S384 (2014).
103. Irwin, M. R. & Cole, S. W. Reciprocal regulation of the neural and innate immune systems. *Nat. Rev. Immunol.* **11**, 625–632 (2011).
104. Lackner, J. M. & Gurtman, M. B. Pain catastrophizing and interpersonal problems: a circumplex analysis of the communal coping model. *Pain* **110**, 597–604 (2004).
105. van Tilburg, M. A., Palsson, O. S. & Whitehead, W. E. Which psychological factors exacerbate irritable bowel syndrome? Development of a comprehensive model. *J. Psychosom Res.* **74**, 486–492 (2013).
106. Camilleri, M. Genetics of human gastrointestinal sensation. *Neurogastroenterol. Motil.* **25**, 458–466 (2013).
107. Saito, Y. A., Mitra, N. & Mayer, E. A. Genetic approaches to functional gastrointestinal disorders. *Gastroenterology* **138**, 1276–1285 (2010).
108. Gupta, A. *et al.* Early life adversity: interactions with stress related gene polymorphisms to impact regional brain structure in females. *Brain Struct. Funct.* http://dx.doi.org/10.1007/s00429-015-0996-9.
109. Kilpatrick, L., Gupta, A., Heendeniya, N., Labus, J. & Mayer, E. A. Corticotropin releasing hormone receptor 1 (crh-r1) and progestoerone receptor (pgr) polymorphisms interact with early life trauma in healthy controls (hc) and patients with irritable bowel syndrome (IBS) [abstract 666]. *Gastroenterology* **144** (Suppl. 1), S121 (2013).
110. Gupta, A. *et al.* Catecholaminergic gene polymorphisms are associated with GI symptoms and morphological brain changes in irritable bowel syndrome. *PLoS ONE* (*in press*).
111. Hong, S., Zheng, G. & Wiley, J. W. Epigenetic regulation of genes that modulate chronic stress-induced visceral pain in the peripheral nervous system. *Gastroenterology* **148**, 148–157.e7 (2015).
112. Tran, L., Schulkin, J., Ligon, C. O. & Greenwood-Van Meerveld, B. Epigenetic modulation of chronic anxiety and pain by histone deacetylation. *Mol. Psychiatry* http://dx.doi.org/10.1038/mp.2014.122.
113. Videlock, E. J. *et al.* Childhood trauma is associated with hypothalamic-pituitary-adrenal axis responsiveness in irritable bowel syndrome. *Gastroenterology* **137**, 1954–1962 (2009).
114. Heim, C. & Nemeroff, C. B. The role of childhood trauma in the neurobiology of mood and anxiety disorders: preclinical and clinical studies. *Biol. Psychiatry* **49**, 1023–1039 (2001).
115. Klengel, T. & Binder, E. B. Gene-environment interactions in major depressive disorder. *Can. J. Psychiatry* **58**, 76–83 (2013).
116. Felitti, V. J. *et al.* Relationship of childhood abuse and household dysfunction to many of the leading causes of death in adults. The Adverse Childhood Experiences (ACE) *Study*. *Am. J. Prev. Med.* **14**, 245–258 (1998).
117. Berman, S. *et al.* Evidence for alterations in central noradrenergic signaling in irritable bowel syndrome. *Neuroimage* **63**, 1854–1863 (2012).
118. Chaloner, A., Rao, A., Al-Chaer, E. D. & Greenwood-Van Meerveld, B. Importance of neural mechanisms in colonic mucosal and muscular dysfunction in adult rats following neonatal colonic irritation. *Int. J. Dev. Neurosci.* **28**, 99–103 (2010).
119. Flugge, G., van Kampen, M., Meyer, H. & Fuchs, E. α2A and α2C-adrenoceptor regulation in the brain: α2A changes persist after chronic stress. *Eur. J. Neurosci.* **17**, 917–928 (2003).
120. Chaloner, A. & Greenwood-Van Meerveld, B. Early life adversity as a risk factor for visceral pain in later life: importance of sex differences. *Front. Neurosci.* **7**, 13 (2013).
121. McEwen, B. S. Understanding the potency of stressful early life experiences on brain and body function. *Metabolism* **57** (Suppl. 2), S11–S15 (2008).
122. Provencal, N. *et al.* The signature of maternal rearing in the methylome in rhesus macaque prefrontal cortex and T cells. *J. Neurosci.* **32**, 15626–15642 (2012).
123. Provencal, N. & Binder, E. B. The effects of early life stress on the epigenome: from the womb to adulthood and even before. *Exp. Neurol.* **268**, 10–20 (2015).
124. Barsky, A. J. & Borus, J. F. Functional somatic syndromes. *Ann. Intern. Med.* **130**, 910–921 (1999).
125. Ahmad, O. F. & Akbar, A. Dietary treatment of irritable bowel syndrome. *Br. Med. Bull.* **113**, 83–90 (2015).
126. Fasano, A., Sapone, A., Zevallos, V. & Schuppan, D. Non-celiac gluten sensitivity. *Gastroenterology* **148**, 1195–1204 (2015).
127. Lomer, M. C. Review article: the aetiology, diagnosis, mechanisms and clinical evidence for food intolerance. *Aliment Pharmacol. Ther.* **41**, 262–275 (2015).
128. Shepherd, S. J., Halmos, E. & Glance, S. The role of FODMAPs in irritable bowel syndrome. *Curr. Opin. Clin. Nutr. Metab. Care* **17**, 605–609 (2014).
129. Grover, M., Camilleri, M., Smith, K., Linden, D. R. & Farrugia, G. On the fiftieth anniversary. Postinfectious irritable bowel syndrome: mechanisms related to pathogens. *Neurogastroenterol. Motil.* **26**, 156–167 (2014).
130. Spiller, R. & Garsed, K. Postinfectious irritable bowel syndrome. *Gastroenterology* **136**, 1979–1988 (2009).
131. Maxwell, P. R., Rink, E., Kumar, D. & Mendall, M. A. Antibiotics increase functional abdominal symptoms. *Am. J. Gastroenterol.* **97**, 104–108 (2002).
132. Villarreal, A. A., Aberger, F. J., Benrud, R. & Gundrum, J. D. Use of broad-spectrum antibiotics and the development of irritable bowel syndrome. *WMJ* **111**, 17–20 (2012).
133. Bashashati, M. *et al.* Cytokine imbalance in irritable bowel syndrome: a systematic review and meta-analysis. *Neurogastroenterol. Motil.* **26**, 1036–1048 (2014).
134. Darkoh, C. *et al.* Chemotactic chemokines are important in the pathogenesis of irritable bowel syndrome. *PLoS ONE* **9**, e93144 (2014).
135. Ohman, L. *et al.* Increased TLR2 expression on blood monocytes in irritable bowel syndrome patients. *Eur. J. Gastroenterol. Hepatol.* **24**, 398–405 (2012).
136. Matricon, J. *et al.* Review article: associations between immune activation, intestinal permeability and the irritable bowel syndrome. *Aliment. Pharmacol. Ther.* **36**, 1009–1031 (2012).
137. Cremon, C. *et al.* Mucosal immune activation in irritable bowel syndrome: gender-dependence and association with digestive symptoms. *Am. J. Gastroenterol.* **104**, 392–400 (2009).
138. Lee, K. J. *et al.* The alteration of enterochromaffin cell, mast cell, and lamina propria T lymphocyte numbers in irritable bowel syndrome and its relationship with psychological factors. *J. Gastroenterol. Hepatol.* **23**, 1689–1694 (2008).
139. Farhadi, A., Fields, J. Z. & Keshavarzian, A. Mucosal mast cells are pivotal elements in inflammatory bowel disease that connect the dots: stress, intestinal hyperpermeability and inflammation. *World J. Gastroenterol.* **13**, 3027–3030 (2007).
140. Cole, S. W. Social regulation of human gene expression. *Curr. Dir. Psychol. Sci.* **18**, 132–137 (2009).
141. Miller, G. E. *et al.* Low early-life social class leaves a biological residue manifested by decreased glucocorticoid and increased proinflammatory signaling. *Proc. Natl Acad. Sci. USA* **106**, 14716–14721 (2009).
142. Szyf, M. The early life environment and the epigenome. *Biochim. Biophys. Acta* **1790**, 878–885 (2009).

© 2015 Macmillan Publishers Limited. All rights reserved

143. Gupta, A. *et al.* Early adverse life events: influence on resting state connectivity in patients with chronic abdominal pain. *Psychosom. Med.* **76**, 404–412 (2014).
144. Cole, S. W. Human social genomics. *PLoS Genet.* **10**, e1004601 (2014).
145. Cole, S. W. *et al.* Transcriptional modulation of the developing immune system by early life social adversity. *Proc. Natl Acad. Sci. USA* **109**, 20578–20583 (2012).
146. Powell, N. D. *et al.* Social stress up-regulates inflammatory gene expression in the leukocyte transcriptome via beta-adrenergic induction of myelopoiesis. *Proc. Natl Acad. Sci. USA* **110**, 16574–16579 (2013).
147. Wohleb, E. S., Powell, N. D., Godbout, J. P. & Sheridan, J. F. Stress-induced recruitment of bone marrow-derived monocytes to the brain promotes anxiety-like behavior. *J. Neurosci.* **33**, 13820–13833 (2013).
148. Wohleb, E. S. *et al.* Re-establishment of anxiety in stress-sensitized mice is caused by monocyte trafficking from the spleen to the brain. *Biol. Psychiatry* **75**, 970–981 (2014).
149. Wohleb, E. S. *et al.* beta-Adrenergic receptor antagonism prevents anxiety-like behavior and microglial reactivity induced by repeated social defeat. *J. Neurosci.* **31**, 6277–6288 (2011).
150. Tillisch, K. *et al.* Sex specific alterations in autonomic function among patients with irritable bowel syndrome. *Gut* **54**, 1396–1401 (2005).
151. Berman, S. *et al.* Differential noradrenergic (NE) modulation of brain metabolic and electrical activity in IBS patients and healthy controls. *Gastroenterology* **132**, A726 (2007).
152. Labus, J. *et al.* Sex-specific differences in a brain network functioning during anticipation of rectal discomfort in irritable bowel syndrome patients (IBS). *Gastroenterology* **130**, A93 (2006).
153. Mayer, E. A. *et al.* Correlation between gene expression profiles in peripheral blood mononuclear cells and structural and functional brain networks in chronic visceral pain. *Neuropsychopharmacology* **39**, S641 (2014).
154. Cole, S. W., Hawkley, L. C., Arevalo, J. M. & Cacioppo, J. T. Transcript origin analysis identifies antigen-presenting cells as primary targets of socially regulated gene expression in leukocytes. *Proc. Natl Acad. Sci. USA* **108**, 3080–3085 (2011).
155. Beissner, F., Meissner, K., Bar, K. J. & Napadow, V. The autonomic brain: an activation likelihood estimation meta-analysis for central processing of autonomic function. *J. Neurosci.* **33**, 10503–10511 (2013).
156. Seminowicz, D. A. *et al.* Regional gray matter density changes in brains of patients with irritable bowel syndrome. *Gastroenterology* **139**, U48–U82 (2010).
157. Cryan, J. F. & Dinan, T. G. Mind-altering microorganisms: the impact of the gut microbiota on brain and behaviour. *Nat. Rev. Neurosci.* **13**, 701–712 (2012).
158. Mayer, E. A. & Tillisch, K. Gut brain axis and the microbiome. *J. Clin. Invest.* **125**, 926–938 (2015).
159. Stilling, R. M., Dinan, T. G. & Cryan, J. F. Microbial genes, brain & behaviour—epigenetic regulation of the gut-brain axis. *Genes Brain Behav.* **13**, 69–86 (2014).
160. Rhee, S. H., Pothoulakis, C. & Mayer, E. A. Principles and clinical implications of the brain-gut-enteric microbiota axis. *Nat. Rev. Gastroenterol. Hepatol.* **6**, 306–314 (2009).
161. Benton, D., Williams, C. & Brown, A. Impact of consuming a milk drink containing a probiotic on mood and cognition. *Eur. J. Clin. Nutr.* **61**, 355–361 (2007).
162. Messaoudi, M. *et al.* Beneficial psychological effects of a probiotic formulation (*Lactobacillus helveticus* R0052 and *Bifidobacterium longum* R0175) in healthy human volunteers. *Gut Microbes* **2**, 256–261 (2011).
163. McNulty, N. P. *et al.* The impact of a consortium of fermented milk strains on the gut microbiome of gnotobiotic mice and monozygotic twins. *Sci. Transl. Med.* **3**, 106ra106 (2011).
164. Mayer, E. A. *et al.* Differences in brain responses to visceral pain between patients with irritable bowel syndrome and ulcerative colitis. *Pain* **115**, 398–409 (2005).
165. Hood, L., Heath, J. R., Phelps, M. E. & Lin, B. Systems biology and new technologies enable predictive and preventative medicine. *Science* **306**, 640–643 (2004).
166. Hornberg, J. J., Bruggeman, F. J., Westerhoff, H. V. & Lankelma, J. Cancer: a systems biology disease. *Biosystems* **83**, 81–90 (2006).
167. Likic, V. A., McConville, M. J., Lithgow, T. & Bacic, A. Systems biology: the next frontier for bioinformatics. *Adv. Bioinformatics* **2010**, 268925 (2010).
168. Chang, L., Lembo, A. & Sultan, S. American Gastroenterological Association institute technical review on the pharmacological management of irritable bowel syndrome. *Gastroenterology* **147**, 1149–1172.e2 (2014).
169. Mayer, E. A. & Collins, S. M. Evolving pathophysiologic models of functional gastrointestinal disorders. *Gastroenterology* **122**, 2032–2048 (2002).
170. Tache, Y., Martinez, V., Million, M. & Wang, L. Stress and the gastrointestinal tract III. Stress-related alterations of gut motor function: role of brain corticotropin-releasing factor receptors. *Am. J. Physiol. Gastrointest. Liver Physiol.* **280**, G173–177 (2001).
171. Patel, V. R., Eckel-Mahan, K., Sassone-Corsi, P. & Baldi, P. CircadiOmics: integrating circadian genomics, transcriptomics, proteomics and metabolomics. *Nat, Methods* **9**, 772–773 (2012).
172. Patel, V. R., Eckel-Mahan, K., Sassone-Corsi, P. & Baldi, P. How pervasive are circadian oscillations? *Trends Cell Biol.* **24**, 329–331 (2014).
173. Partch, C. L., Green, C. B. & Takahashi, J. S. Molecular architecture of the mammalian circadian clock. *Trends Cell Biol.* **24**, 90–99 (2014).
174. Eckel-Mahan, K. L. *et al.* Reprogramming of the circadian clock by nutritional challenge. *Cell* **155**, 1464–1478 (2013).
175. Takahashi, J. S., Hong, H. K., Ko, C. H. & McDearmon, E. L. The genetics of mammalian circadian order and disorder: implications for physiology and disease. *Nat. Rev. Genet.* **9**, 764–775 (2008).
176. Bellet, M. M. *et al.* Circadian clock regulates the host response to Salmonella. *Proc. Natl Acad. Sci. USA* **110**, 9897–9902 (2013).
177. Zhang, R., Lahens, N. F., Ballance, H. I., Hughes, M. E. & Hogenesch, J. B. A circadian gene expression atlas in mammals: implications for biology and medicine. *Proc. Natl Acad. Sci. USA* **111**, 16219–16224 (2014).
178. Bonaz, B. L. & Bernstein, C. N. Brain-gut interactions in inflammatory bowel disease. *Gastroenterology* **144**, 36–49 (2013).
179. Greicius, M. D., Krasnow, B., Reiss, A. L. & Menon, V. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proc. Natl Acad. Sci. USA* **100**, 253–258 (2003).
180. Buckner, R. L., Andrews-Hanna, J. R. & Schacter, D. L. The brain's default network: anatomy, function, and relevance to disease. *Ann. N. Y. Acad. Sci.* **1124**, 1–38 (2008).
181. Raichle, M. E. *et al.* A default mode of brain function. *Proc. Natl Acad. Sci. USA* **98**, 676–682 (2001).
182. Qin, P. & Northoff, G. How is our self related to midline regions and the default-mode network? *Neuroimage* **57**, 1221–1233 (2011).
183. Damoiseaux, J. S. *et al.* Consistent resting-state networks across healthy subjects. *Proc. Natl Acad. Sci. USA* **103**, 13848–13853 (2006).
184. Labus, J. S. *et al.* Sex differences in brain activity during aversive visceral stimulation and its expectation in patients with chronic abdominal pain: a network analysis. *Neuroimage* **41**, 1032–1043 (2008).
185. Pezawas, L. *et al.* 5-HTTLPR polymorphism impacts human cingulate-amygdala interactions: a genetic susceptibility mechanism for depression. *Nat. Neurosci.* **8**, 828–834 (2005).
186. Stein, J. L. *et al.* A validated network of effective amygdala connectivity. *Neuroimage* **36**, 736–745 (2007).
187. Critchley, H. D., Nagai, Y., Gray, M. A. & Mathias, C. J. Dissecting axes of autonomic control in humans: Insights from neuroimaging. *Auton. Neurosci.* **161**, 34–42 (2011).
188. Saper, C. B. The central autonomic nervous system: conscious visceral perception and autonomic pattern generation. *Annu. Rev. Neurosci.* **25**, 433–469 (2002).
189. Biswal, B., Yetkin, F. Z., Haughton, V. M. & Hyde, J. S. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magn. Reson. Med.* **34**, 537–541 (1995).
190. Jones, L. M., Fontanini, A., Sadacca, B. F., Miller, P. & Katz, D. B. Natural stimuli evoke dynamic sequences of states in sensory cortical ensembles. *Proc. Natl Acad. Sci. USA* **104**, 18772–18777 (2007).
191. Alcauter, S. *et al.* Development of thalamocortical connectivity during infancy and its cognitive correlations. *J. Neurosci.* **34**, 9067–9075 (2014).
192. Levinthal, D. J. & Strick, P. L. The motor cortex communicates with the kidney. *J. Neurosci.* **32**, 6726–6731 (2012).
193. Menon, V. Large-scale brain networks and psychopathology: a unifying triple network model. *Trends Cogn. Sci.* **15**, 483–506 (2011).
194. Borsook, D., Edwards, R., Elman, I., Becerra, L. & Levine, J. Pain and analgesia: the value of salience circuits. *Prog. Neurobiol.* **104**, 93–105 (2013).
195. Greicius, M. D. *et al.* Resting-state functional connectivity in major depression: abnormally increased contributions from subgenual cingulate cortex and thalamus. *Biol. Psychiatry* **62**, 429–437 (2007).
196. Laird, A. R. *et al.* Behavioral interpretations of intrinsic connectivity networks. *J. Cogn. Neurosci.* **23**, 4022–4037 (2011).
197. Grupe, D. W. & Nitschke, J. B. Uncertainty and anticipation in anxiety: an integrated neurobiological and psychological perspective. *Nat. Rev. Neurosci.* **14**, 488–501 (2013).

Acknowledgements

Supported by funding National Institutes of Health grants P50 DK064539 (E.A.M.), R01 DK048351 (E.A.M.) and P30 DK041301 (E.R.). The authors thank Cathy Liu and Arpana Gupta for invaluable editorial assistance.

Author contributions

The article was conceptualized and composed by E.A.M. Inputs to the following sections came from the following co-authors: immune system (S.W.C.); brain networks (J.S.L.); gut microbiota (K.T.); and systems biology (P.B.).

© 2015 Macmillan Publishers Limited. All rights reserved

Irritable bowel syndrome

Paul Enck[1], Qasim Aziz[2], Giovanni Barbara[3], Adam D. Farmer[2], Shin Fukudo[4], Emeran A. Mayer[5], Beate Niesler[6], Eamonn M. M. Quigley[7], Mirjana Rajilić-Stojanović[8], Michael Schemann[9], Juliane Schwille-Kiuntke[1], Magnus Simren[10], Stephan Zipfel[1] and Robin C. Spiller[11]

Abstract | Irritable bowel syndrome (IBS) is a functional gastrointestinal disease with a high population prevalence. The disorder can be debilitating in some patients, whereas others may have mild or moderate symptoms. The most important single risk factors are female sex, younger age and preceding gastrointestinal infections. Clinical symptoms of IBS include abdominal pain or discomfort, stool irregularities and bloating, as well as other somatic, visceral and psychiatric comorbidities. Currently, the diagnosis of IBS is based on symptoms and the exclusion of other organic diseases, and therapy includes drug treatment of the predominant symptoms, nutrition and psychotherapy. Although the underlying pathogenesis is far from understood, aetiological factors include increased epithelial hyperpermeability, dysbiosis, inflammation, visceral hypersensitivity, epigenetics and genetics, and altered brain–gut interactions. IBS considerably affects quality of life and imposes a profound burden on patients, physicians and the health-care system. The past decade has seen remarkable progress in our understanding of functional bowel disorders such as IBS that will be summarized in this Primer.

Irritable bowel syndrome (IBS) is a functional bowel disorder (that is, not associated with structural or biochemical abnormalities that are detectable with the current routine diagnostic tools) characterized by abdominal pain or discomfort, stool irregularities and bloating (BOX 1). Symptoms can be debilitating in many individuals, but may be mild or moderate in other patients. In addition, IBS is often associated with other somatic comorbidities (for example, pain syndromes, overactive bladder and migraine), psychiatric conditions (including depression and anxiety) and visceral sensitivity. The population prevalence of IBS is high (~11%) and the condition has considerable consequences for quality of life (QOL) that are comparable to other chronic diseases, such as diabetes mellitus and hepatitis. IBS is diagnosed based on symptoms, and a distinction is made between the following subtypes of IBS: IBS with pain or discomfort and predominant constipation (IBS-C), IBS with diarrhoea (IBS-D), mixed IBS (IBS-M) and unsubtyped IBS (IBS-U) (FIG. 1). Moreover, other diseases (including other functional gastrointestinal diseases, such as functional dyspepsia and gastroesophageal reflux disease) that may cause the typical IBS symptoms should be excluded. Although a substantial proportion of patients will experience spontaneous remission over time, there is currently no treatment that cures IBS; relief of symptoms is the most that can be achieved.

IBS is a multifactorial disease. Hence, the underlying pathogenesis is considered complex and the precise molecular pathophysiology is far from understood. Several functional alterations have been described, such as altered visceral sensitivity, functional brain alterations, bowel motility and secretory dysfunctions, and somatic and psychiatric comorbidities. Furthermore, gastrointestinal abnormalities — such as immune activation, gut dysbiosis (microbial imbalance), impaired mucosal functions, nerve sensitization, post-infectious plasticity, altered expression and release of mucosal and immune mediators, and altered gene expression profiles — have been associated with IBS. However, a coherent link between particular pathologies and IBS symptoms is yet to be established.

Moreover, results from studies assessing the contribution of most of the proposed pathological factors are inconsistent and the particular aetiology is often not related to particular gut symptoms. For example, some studies have found evidence for gut micro-inflammation in IBS, whereas others could not confirm this finding, despite similar gastrointestinal symptoms. Such discrepancies, which also apply to the other biomarker candidates (not only to inflammation), strongly suggest the existence of IBS subpopulations, which, despite the similarity in gut symptoms, can be defined and distinguished by their pathophysiology and in-depth

Correspondence to P.E.
Department of Internal Medicine VI (Psychosomatic Medicine and Psychotherapy), University Hospital Tübingen, Osianderstraße 5, 72076 Tübingen, Germany.
paul.enck@uni-tuebingen.de

Article number: 16014
doi:10.1038/nrdp.2016.14
Published online 24 March 2016

© 2016 Macmillan Publishers Limited. All rights reserved.

Author addresses

[1]Department of Internal Medicine VI (Psychosomatic Medicine and Psychotherapy), University Hospital Tübingen, Tübingen, Germany.
[2]Wingate Institute of Neurogastroenterology, Barts and London School of Medicine and Dentistry, Queen Mary University of London, London, UK.
[3]Department of Medical and Surgical Sciences, St. Orsola-Malpighi Hospital, Bologna, Italy.
[4]Department of Behavioural Medicine, Tohoku University Graduate School of Medicine, Sendai, Japan.
[5]Oppenheimer Center for Neurobiology of Stress, Division of Digestive Diseases, David Geffen School of Medicine at UCLA, Los Angeles, California, USA.
[6]Department of Human Molecular Genetics, University of Heidelberg, Heidelberg, Germany.
[7]Lynda K and David M Underwood Center for Digestive Disorders, Division of Gastroenterology and Hepatology, Houston Methodist Hospital, Weill Cornell Medical College, Houston, Texas, USA.
[8]Department of Biochemical Engineering and Biotechnology, Faculty of Technology and Metallurgy, University of Belgrade, Belgrade, Serbia.
[9]Department of Human Biology, Technical University Munich, Freising-Weihenstephan, Germany.
[10]Department of Internal Medicine and Clinical Nutrition, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden.
[11]NIHR Nottingham Digestive Diseases Biomedical Research Unit, University of Nottingham, Nottingham, UK.

assessments of clinical and molecular biomarker clusters. The same heterogeneity is evident with respect to clinical diagnosis and management. Indeed, medical treatment, nutritional intervention and psychotherapy lack consistent and homogeneous efficacy, but can be effective in some subgroups.

This Primer summarizes recent progress in our understanding of IBS prevalence, comorbidities, QOL and the putative roles of inflammation, genetics, the intestinal microbiota and the brain–gut axis in IBS pathogenesis. Furthermore, we will discuss the current diagnostic approach and highlight the therapeutic options in IBS, including drugs, nutrition and psychotherapy.

Epidemiology

Global prevalence and incidence

Prevalence rates of IBS vary between 1.1% and 45%, based on population studies from countries worldwide (FIG. 2; Supplementary information S1 (table)), with a pooled global prevalence of 11.2% (95% CI: 9.8–12.8)[1]. Prevalence rates of 5–10% are reported for most European countries, the United States and China[1]. Population statistics for IBS in most African and many Asian countries are unavailable, which might point to the inability to differentiate between infectious diarrhoea and IBS in tropical countries, especially in those nations with poor health-care systems or limited patient access to medical care, or to less attention of the health-care system for functional disorders, once an acute infection has been excluded[2].

Gathering subtype-specific prevalence information is complex. IBS subtypes overlap considerably in terms of symptoms, and patients vary over time in terms of their predominant symptoms, and thus switch subtype[3]. The few population studies that have differentiated between IBS subtypes suggest that, in countries with a total IBS prevalence of ~10%, IBS-C and IBS-D each account for one-third of the affected population[4]. Incidence rates of IBS (that is, the annual occurrence of new cases) are not reported for most countries, but a few long-term surveys (≥10 years) in the United States allow for an estimation of the annual incidence in the range of 1–2%[5]. At the same time, disappearance rates of 2% have been reported[6], indicating spontaneous disease remission.

Association between IBS and other disorders

Not only do IBS subtypes overlap[6] but population-based studies also report a substantial overlap of ≥20% with other functional gastrointestinal disorders of the upper and lower gastrointestinal system: functional dyspepsia, heartburn, gastroesophageal reflux disease and nausea on the one hand[7], and diarrhoea, incontinence, pelvic floor dyssynergia and constipation on the other hand[8]. An overlap of IBS with inflammatory bowel diseases (IBDs; including Crohn disease and ulcerative colitis) during remission phases has been proposed[9] but is not mutually agreed on[10].

Other IBS-associated disorders (FIG. 3) include functional non-gastrointestinal syndromes, such as urological chronic pelvic pain syndrome (this term includes interstitial cystitis and chronic prostatitis), vulvodynia, overactive bladder, prostatic pain syndrome, premenstrual syndrome, sexual (including erectile) dysfunction, chronic pelvic pain, fibromyalgia syndrome, chronic fatigue syndrome, migraine, eating disorders, nutritional intolerances and others[11]. All of these syndromes considerably overlap with IBS in population studies to a degree that is often beyond what is expected based on the prevalence rates of the individual diseases. Given that many of these conditions are only diagnosed in specialized centres, it has been questioned as to whether some of these conditions — for example, IBS and chronic pelvic pain — are one and the same disease[12].

In addition, most epidemiological studies note the presence of psychiatric comorbidities (such as anxiety, depression, somatization or neuroticism) not only for IBS but also for these IBS-associated diseases. Again, the rates are above the expected levels for IBS and the population prevalence of these symptoms[13]. Thus, the entire disease entity (IBS, functional gastrointestinal disorders and other functional non-gastrointestinal disorders) has been included in the term 'somatic symptom disorder'

Box 1 | **IBS definition and subtypes: Rome III criteria**

Diagnostic criteria* for irritable bowel syndrome (IBS) include recurrent abdominal pain or discomfort‡ at least 3 days per month in the past 3 months associated with two or more of the following:

- Improvement with defaecation
- Onset associated with a change in the frequency of stool
- Onset associated with a change in the form (appearance) of stool

*Criteria fulfilled for the past 3 months with symptom onset at least 6 months before diagnosis.
‡Discomfort means an uncomfortable sensation not described as pain. In pathophysiological research and clinical trials, a pain or discomfort frequency of at least 2 days per week during screening evaluation for subject eligibility. Adapted with permission from REF. 119, American Gastroenterology Association.

© 2016 Macmillan Publishers Limited. All rights reserved.

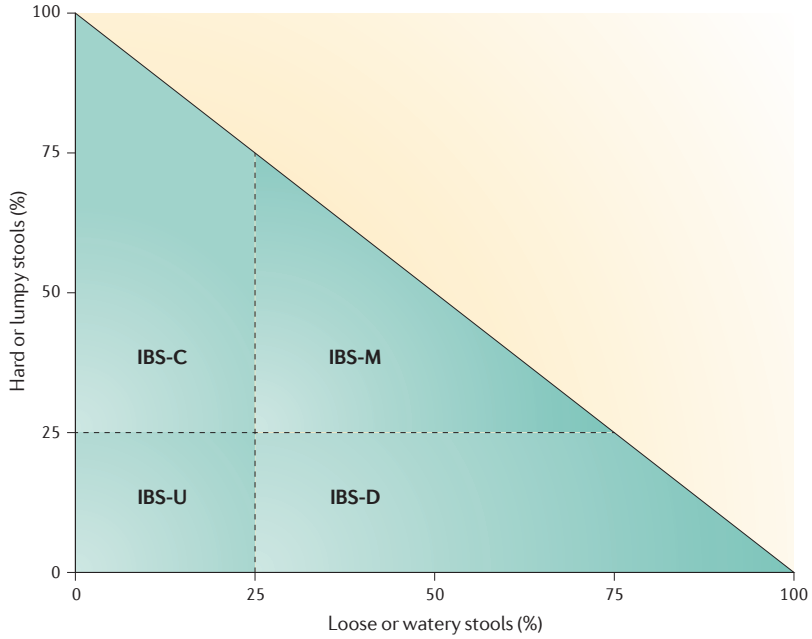


Figure 1 | **IBS subtypes according to the Rome III criteria.** A two-dimensional graph of the four possible irritable bowel syndrome (IBS) subtypes according to bowel form at a particular point in time, and the percentage of time this bowel form has to be present to meet the criteria for IBS with constipation (IBS-C), IBS with diarrhoea (IBS-D), mixed-type IBS (IBS-M) and unsubtyped IBS (IBS-U). Adapted with permission from REF. 119, American Gastroenterology Association.

in the Diagnostic and Statistical Manual of Mental Disorders 5th Edition (DSM-5)[14] and in psychiatric or psychosomatic clinical management[15]. Patients with IBS who were treated by psychiatrists frequently did not receive adequate attention with respect to their gastrointestinal symptoms before the release of DSM-5.

Risk factors for IBS

The best-documented risk factor for IBS is female sex, which is associated with an odds ratio of 1.67 (95% CI: 1.53–1.82) across many population-based studies[16], with explanations varying between sex-different health care, consultation behaviour and biological functions (for example, hormonal regulation of gut functions). The incidence of IBS decreases with advancing age (>50 years)[1], but is similar in children and adolescents compared with adults and does not necessarily transmit from childhood to adulthood[17]. However, family aggregation has been reported[18] that is driven by genetics[19] as well as by social learning[20]. BOX 2 lists the personal, disease, psychosocial and social factors that have been found to be associated with increased risk of IBS, although some of these factors have only been identified in individual studies[21] or have been found to vary between countries and settings.

Post-infectious IBS

Several studies have shown an association between IBS and preceding gastrointestinal infections of bacterial, viral or other origin[22,23]. The pooled odds ratio is 7.3 (95% CI: 4.7–11.1) for the development of IBS after infectious gastroenteritis[24], with a median prevalence of ~10%[22]. This association seems to differ with respect to epidemic infectious events that affect many people at the same time and individual infections, such as travellers' diarrhoea. That is, prevalence data are reported to be higher (15–30%) in epidemic events[22] and lower (5–10%) following travellers' diarrhoea[23]; these differences are presumably due to different reporting biases in these populations. Thus, a median prevalence of 10% might better reflect the true prevalence of post-infectious IBS than the extreme values reported in individual studies. Risk factors for the development of post-infectious IBS are female sex, younger age, the severity of the initial infection and premorbid psychological conditions[22–24]. Based on symptoms alone, post-infectious IBS cannot be distinguished from IBS without an infectious origin, but inflammatory biomarkers may. The most valid distinction may be a sudden onset that is well remembered by the patient and is associated with fever, bloody stools and a positive laboratory stool test for an infective agent.

Mechanisms/pathophysiology

Although the aetiology of IBS remains largely undetermined, our understanding of the potential mechanisms involved in gut dysfunction, visceral sensation and symptom generation is rapidly advancing. Growing evidence suggests that, in IBS, the epithelial barrier, gut microbiota, food antigens and bile acids elicit abnormal responses in the key regulators of sensorimotor functions, including the hypothalamus–pituitary–adrenal (HPA) axis, the immune system, the brain–gut axis and the enteric nervous system (ENS) (FIG. 4). Accordingly, these factors might have a role as potential biomarkers of disease (BOX 3). In addition to these putative biomarkers, psychological factors ('psychomarkers') such as depression and anxiety, which are known to respond to abdominal symptoms (bottom-up), and psychosocial factors ('stress') that influence physiological (intestinal) functions, such as motility and visceral sensitivity (top-down), have been acknowledged and will be discussed in more detail.

The epithelial barrier

The epithelial gut lining represents an enormous surface that is in constant contact with the environment and with billions of bacteria that constantly challenge the intestinal immune system. Increased intestinal permeability is considered an early event in IBS that leads to low-grade immune cell infiltration of the gut mucosa[25]. Indeed, increased epithelial permeability has been primarily described in post-infectious IBS in general and in IBS-D in particular, although some reports have also shown that IBS-C and IBS-M might also involve an increase in epithelial permeability[25]. Evidence for the presence of this remodelling in IBS has been provided by electron microscopy, which has detected enlarged spaces between epithelial cells and cytoskeletal condensation in gut biopsies of patients with IBS-D[26]. In addition, Ussing chamber experiments, which measure epithelial membrane properties on colonic mucosal biopsies, have shown excessive passage of macromolecules from the luminal to the basolateral side of gut tissue

© 2016 Macmillan Publishers Limited. All rights reserved.

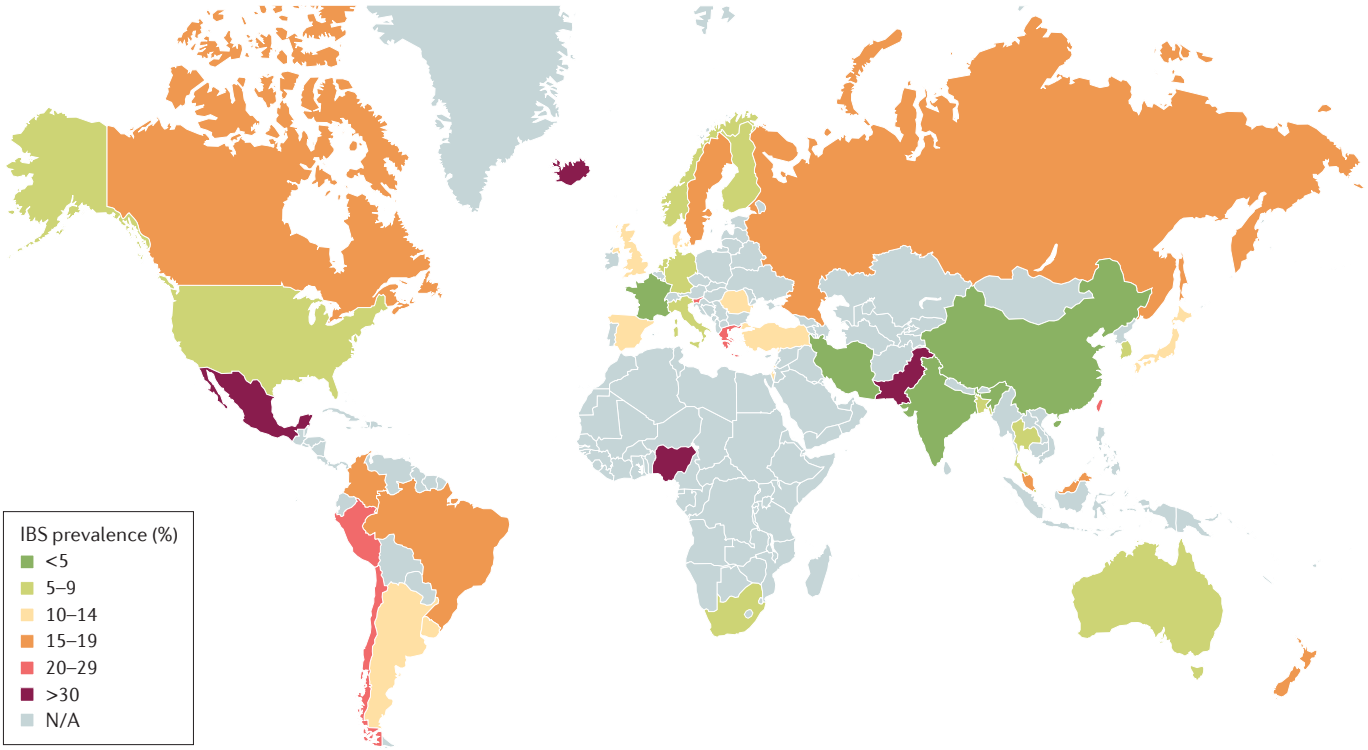


Figure 2 | **IBS prevalence in population studies around the world.** Pooled prevalence data per country are colour-coded. Data from REF. 1 are supplemented by studies from another nine countries (see Supplementary information S1 (table)). IBS, irritable bowel syndrome; N/A, not applicable.

in biopsies obtained from patients with IBS compared with asymptomatic controls, hence providing the functional correlate for the described structural epithelial barrier defects[27].

Morphological and functional changes in intestinal permeability are related to abnormal gene and protein expression of tight junction proteins, including a reduction in the expression of occludin and zonula occludens protein 1 (REFS 25,28). These findings have recently been corroborated by genetic and epigenetic findings in tight junction proteins claudin 1, claudin 2 and cingulin, as outlined below. Tight junction changes are probably the result of both bacterial-mediated and proteasome-mediated degradation triggered by low-grade inflammation[29]. Accordingly, inflammatory mediators including eicosanoids, histamine and proteases increase intestinal permeability. This may involve the participation of ENS neurons, which may amplify these effects[27,30].

Increased intestinal permeability has been linked to diarrhoea and pain severity[26], suggesting that this mechanism might have a role in symptom generation in IBS. Although the exact causes underlying the 'leaky' gut barrier in IBS remain elusive, it has been postulated that numerous factors could be involved, including genetics, epigenetics, dysbiosis and food allergies[25]. Confocal laser endomicroscopy of the duodenal mucosa of patients with IBS after challenge with food to which the patients reported intolerance showed epithelial breaks and increased intervillous spaces, indicative of increased intestinal permeability. These studies suggest a causative effect of food in the increased epithelial permeability in IBS[31].

Bile acids

A subset of patients with features compatible with IBS-D present with increased levels of total faecal bile acids caused by increased excretion and synthesis of serum C4 (7α-hydroxy-4-cholesten-3-one; a surrogate for bile acid synthesis), which in turn influences bowel habit by accelerating colonic transit and inducing diarrhoea and visceral hypersensitivity in IBS[32–34]. Of note, genes involved in bile acid metabolism and function have been reported to be associated with colonic transit in IBS-D, as outlined below.

Immune response

It has been argued that the immune system participates in the pathophysiology of IBS based on the clinical observation that infectious gastroenteritis is a strong risk factor for the development of IBS[24]. Additional clinical support comes from the evidence that about one-third of patients with IBD in remission experience IBS-like symptoms[35]. These inferential data have been subsequently enriched by quantitative immunohistochemistry data showing increased infiltration of T cells and mast cells in the mucosa of the small and large intestine of some patients with IBS[36].

Two randomized controlled trials (RCTs)[37,38] in patients with IBS demonstrated that the anti-inflammatory agent mesalazine was not superior to placebo in alleviating IBS symptoms, although both studies clearly indicated that subgroups, particularly patients with post-infectious IBS, had sustained symptomatic responses. Thus, these studies confirm the hypothesis that immune activation has a considerable role in some patients with IBS.

© 2016 Macmillan Publishers Limited. All rights reserved.

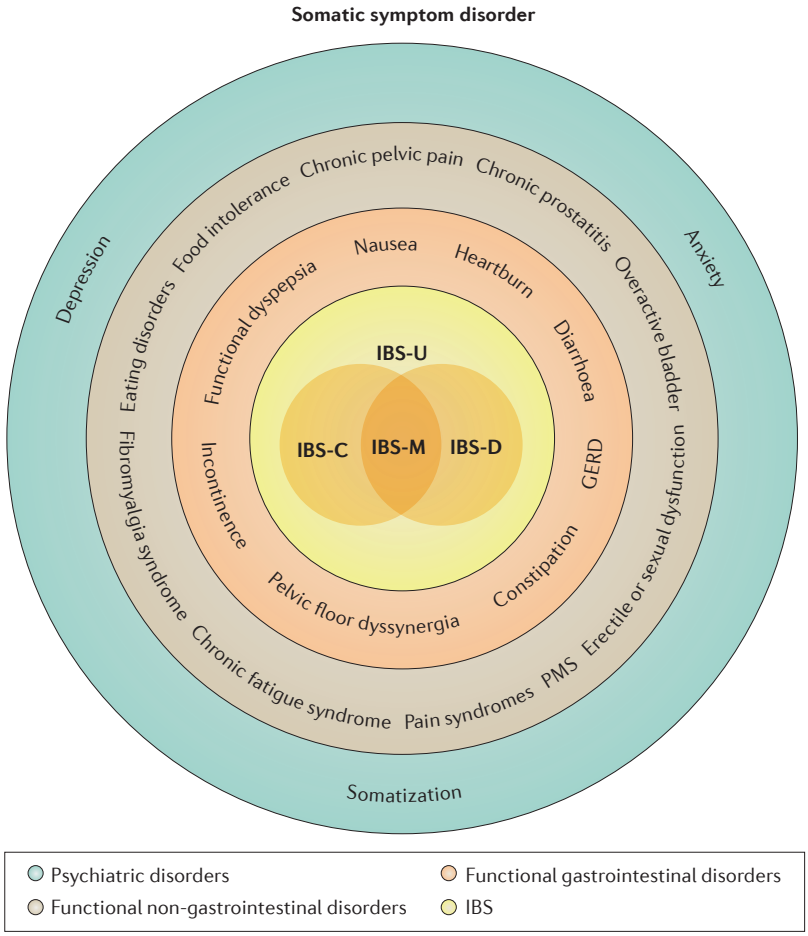


Figure 3 | **IBS-associated comorbidities.** A model of irritable bowel syndrome (IBS) and its associations with other clinical, intestinal, extra-intestinal and psychiatric conditions. For each of the listed disorders, overlap with IBS symptoms has been reported in the literature[11]. The different components should be viewed as layers of complexity: the IBS subtypes are part of the group of functional bowel disorders, these are part of all kinds of functional disorders and these again are part of a 'layer' of psychiatric disorders. GERD, gastroesophageal reflux disease; IBS-C, IBS with constipation; IBS-D, IBS with diarrhoea; IBS-M, mixed-type IBS; IBS-U, unsubtyped IBS; PMS, premenstrual syndrome.

Although mucosal immunocyte numbers are not always increased in IBS, there is strong functional and molecular evidence of an increased state of activation of immune cells in about half of patients with IBS[36]. Data from several studies point to the importance of mast cells as key components of inducing and maintaining low-grade immune activation in IBS[36]. For instance, higher proportions of mast cells were found in a degranulating state in colonic biopsies from patients with IBS than in control samples, suggesting that increased activation of mast cells is involved in the condition[39]. In addition, biopsy supernatants from patients with IBS contained higher amounts of mast cell mediators, including proteases and histamine[36] as well as polyunsaturated fatty acid metabolites[40], than controls. Mucosal immune activation is coupled with altered gene expression of several components of the host mucosal immune response to microbial pathogens (see below), suggesting that the microbiota might contribute to the observed immune activation[36].

Neuroimmune interactions

Mucosal mediators isolated from biopsy samples from patients with IBS have been extensively studied to identify their effect on bowel physiology and sensory perception in isolated tissues or laboratory animals[41]. Compared with controls, mucosal mediators from patients with IBS evoked higher activation of visceral and somatic pain pathways when applied to intestinal preparations isolated from rodents[42,43]. Mast cells and enteroendocrine cells have been suggested to participate in this abnormal neural signalling, as indicated by the activation of human ENS neurons via mast cell-derived histamine, enteroendocrine cell-derived serotonin (also known as 5-hydroxytryptamine (5-HT)) and protease-dependent mechanisms[30,42] (FIG. 5). Although most of the proteases are secreted by mast cells, some of the serine and cysteine proteases that are present at a higher level in the mucosa or stool of patients with IBS than controls might be of other, probably pancreatic or bacterial, origin. In line with these findings, serine proteases in faecal supernatants from individuals with IBS-D evoked colonic hypersensitivity to distension[44]. By contrast, faecal cysteine protease activity was augmented in some patients with IBS-C compared with controls and increased protease activity correlated with abdominal pain and impaired epithelial permeability[45]. Further work showed the implication of serine proteases that act on protease-activated nociceptors located on intestinal nerves conveying pain stimuli to the brain[43]. Importantly, mucosal mediators from patients with IBS and visceral hypersensitivity — but not from normosensitive patients with IBS — acutely activated spinal nociceptors when given to animal models[46]. In the same model, chronic exposure to soluble mediators from patients with IBS-D was shown to sensitize nociceptive neurons[47], implying that chronicity is associated with long-lasting plasticity alterations.

Attention has been directed to agonists of the transient receptor potential cation channels (TRPs), which have been implicated in the pathogenesis of sensory hyperalgesia. Colon tissue samples from patients with IBS have increased levels of specific polyunsaturated fatty acids, which stimulate sensory neurons from mice via the activation of TRP subfamily V member 4 (TRPV4) and generate visceral hypersensitivity[40]. The importance of those visceral afferents that express TRPs in IBS symptomatology is underscored by the finding that peripheral blood mononuclear cell (PBMC) supernatants from patients with IBS-D cause mechanical hypersensitivity of visceral afferents via tumour necrosis factor (TNF) and TRPA1; this was not observed if control supernatants were used[48].

Recent data support the concept that the chronic release of factors with known effects on nerves in the intestinal milieu might not only have functional effects but could also affect the ENS and sensory fibres in a structural manner. For example, immunohistochemistry showed a 57.7% and 56.1% increase in mucosal neurons and neuronal outgrowth, respectively, in patients with IBS compared with healthy controls[49]. Indeed,

© 2016 Macmillan Publishers Limited. All rights reserved.

the intestinal mucosa of patients with IBS contains increased levels of nerve growth factor (NGF), primarily in mucosal mast cells. Experimentally, the effect of NGF was demonstrated in primary cell cultures of the rat myenteric plexus and the neuroblastoma cell line SH-SY5Y, which showed an increase in neurite growth, and protein and mRNA expression of growth-associated protein 43 (GAP43; also known as neuromodulin) — a key neuronal growth protein — following exposure to supernatant obtained from mucosal biopsies of patients with IBS[49].

Box 2 | **Risk factors for IBS**

Personal factors
- Sex (female)
- Age (>50 years)
- Birth cohort*
- Breast feeding (<6 months)*
- Herbivore pet in childhood*
- Birth weight (low)*
- Body mass index (low)*

Psychological factors
- Illness behaviour
- Low quality of life
- Acute psychological stress
- Stressful life events
- Sexual or physical abuse history
- Anxiety, depression or somatization
- Intimate partner violence*
- Addictive behaviour*

Somatic issues
- Gastrointestinal infection
- Somatic symptoms (pains, for example, joint pain and migraine)
- Endometriosis
- Abdominal obesity
- Diverticular disease (left side)
- Antibiotic use
- Abdominal surgery
- Spicy food consumption*
- Sleep problems*
- Low exercise level*

Social conditions
- Socioeconomic status (childhood)
- Family history of substance abuse
- Family history of mental illness
- Working conditions (insufficient autonomy)*
- Shift work*
- Marital status (never married)*
- Number of family members (with more members increasing the risk)*
- Childhood war exposure*

Less well-established factors are marked (*) and are based on single studies (for example, REF. 21), whereas all others have been shown in more than one study.

Microbiota

The gastrointestinal microbiota is a diverse and numerous ecosystem that inhabits the entire gastrointestinal tract and has a systemic influence on our health. Owing to its enormous complexity and high interindividual variability, the microbiota is still in large part undefined regarding the scope of its contribution to human physiology and tolerable compositional variations under which normal functions are preserved[50]. The evidence for an involvement of altered gut microbiota composition in IBS pathophysiology has been accumulating (BOX 4), but the aetiological role remains uncertain. The most prominent markers of IBS are derived from uncultured bacteria. Two groups of uncultured Clostridiales are significantly depleted in IBS[51,52], and bacteria related to *Ruminococcus torques* (a species belonging to the Lachnospiraceae) are profoundly enriched in patients with IBS[51,53,54] and levels positively correlate with bowel symptoms[51,52,55]. In addition, increased Firmicutes to Bacteroidetes ratios have been observed at the phylum level, at least in a subset of patients[51] (for a recent review see REF. 56). Given the provided evidence, the dysbiosis of microbiota in IBS has been acknowledged by the Rome Foundation Working Team[57] as a plausible contributing factor to the disorder. Experiments with animal models have shown that colonization of germ-free animals with microbiota from patients with IBS can induce visceral hypersensitivity[58], impair intestinal permeability and alter gastrointestinal transit time[59] — indicating the importance and the possible aetiological role of the microbiota in IBS.

Although diet changes have an effect on the abundance of particular microbial groups, the microbiotic signature (in terms of present species) is very stable[60]. To observe a profound effect, the dietary change has to be dramatic (for example, vegans switching to high-fat and high-protein diets[61]). Dietary interventions (such as low dietary content of fermentable oligosaccharides, disaccharides, monosaccharides and polyols (FODMAPs; BOX 5), or the addition of sweeteners (fructo-oligosaccharides) or fibre (psyllium)) can improve symptoms of some but not all patients with IBS. Future studies should evaluate the relevance of these microbial groups for IBS and could contribute to a better understanding of the role of the microbiota in the pathophysiology of IBS that is currently acknowledged for the following contexts.

Fermentation of non-digestible foods. An important role of the microbiota is degradation of non-digestible dietary components[62]. It is generally accepted that fermentation of carbohydrates is desirable because of the beneficial effects of the main fermentation products — short-chain fatty acids (SCFAs) — including energy supply to gastrointestinal epithelial cells, a decrease in inflammation and improvement in gut barrier function[63]. However, in patients with IBS, the presence of the resistant carbohydrates FODMAPs can provoke IBS symptoms[64]. This might be a result of overproduction or underproduction of relevant metabolites owing to the disturbed microbiotic balance, for example,

© 2016 Macmillan Publishers Limited. All rights reserved.

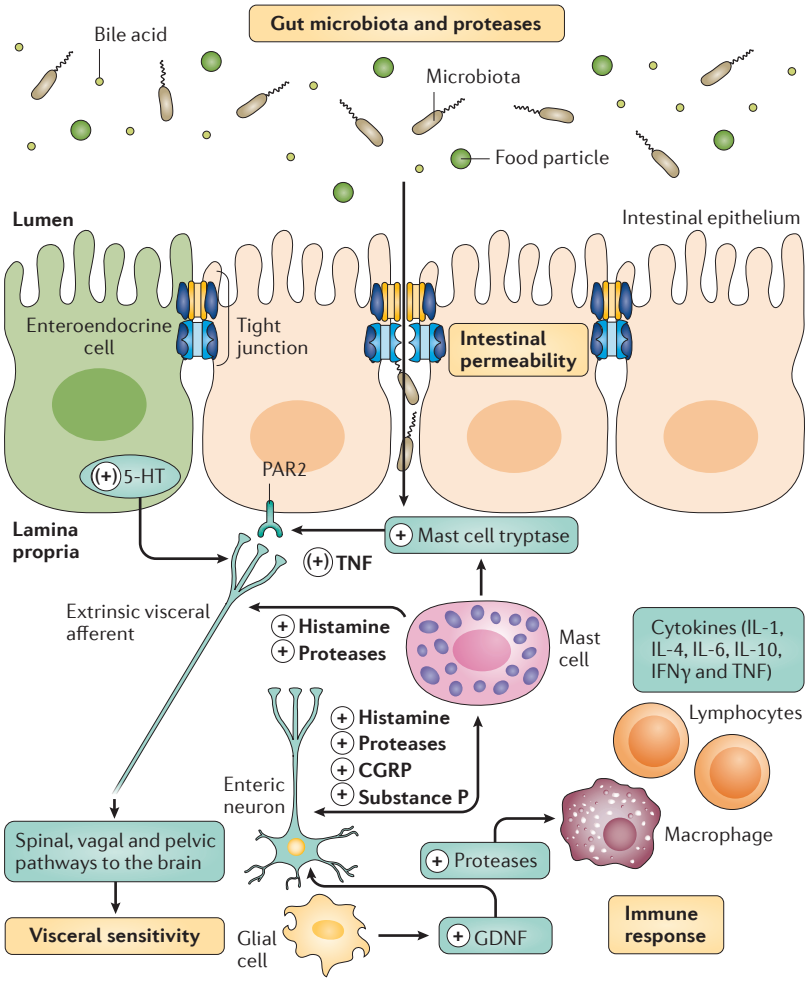


Figure 4 | **Overview of the pathophysiology of IBS.** Although the aetiology of irritable bowel syndrome (IBS) has not yet been completely elucidated, various factors have a role, including composition of the gut microbiota, intestinal permeability, immune cell reactivity and sensitivity of the enteric nervous system, the brain–gut axis (spinal, vagal or pelvic pathways) or the brain. The figure highlights those mediators that are probably involved in IBS pathology. The plus symbols indicate whether a mediator activates or inhibits its target cell; those in parentheses denote actions established in animal models and those without parentheses are effects demonstrated in humans (human tissue). 5-HT, 5-hydroxytryptamine (also known as serotonin); CGRP, calcitonin gene-related peptide; GDNF, glial cell-derived neurotrophic factor; IL, interleukin; PAR2, proteinase-activated receptor 2; TNF, tumour necrosis factor.

due to an increased abundance of gas-producing and decreased abundance of gas-utilizing microorganisms. The quantity and composition of SCFAs in the gut differs between patients with IBS and healthy controls, although the available data are not always in agreement[65,66]. Moreover, the production of microbial SCFAs stimulates regulatory T cell differentiation and affects the balance between pro-inflammatory and anti-inflammatory mechanisms[67], suggesting that inadequate levels of SCFAs could provoke low-grade intestinal inflammation as observed in patients with IBS[68,69]. Finally, studies of microbiota show that the abundance of several SCFA-producing bacteria — including *Roseburia*, *Blautia* and *Veillonella*[70] — is significantly increased compared with the levels of these bacteria in healthy controls, providing a potential mechanistic basis for the development of IBS symptoms.

Other carbohydrate-utilizing gastrointestinal bacteria — namely, *Dorea* spp. — show significant increases in abundance in patients with IBS[51]; these are the main gas-producing bacteria in the human gastrointestinal tract[71]. The overproduction of gas is associated with IBS[72] and this phenomenon could underlie flatulence and abdominal pain. The excessive production of gas can also cause faster colonic transit in patients with IBS-D, as the colons of these patients are more sensitive to increased intestinal volume than healthy controls[73]. Intestinal gases are efficiently removed by methanogenic archaea[74], which seem to be depleted in patients with IBS[51,52] and are negatively correlated with the presence of loose stools[52]. However, a significant increase in the abundance of this microbial group is characteristic of patients with slow transit and constipation[75], whereas the degree of the methanogenic activity could be correlated with the severity of constipation in those with IBS-C[76].

Another potential pathway for microbiotic involvement in IBS is protein degradation. The luminal contents of patients with IBS contain increased levels of proteases[30], which could be due to the increased secretion of endogenous and microbial proteases in response to protein-rich nutrition (typical of western diets), but could also be due to insufficient endogenous protease degradation by the disturbed gastrointestinal microbial community[77]. Serine protease inhibitors are produced by many bacteria, including bifidobacteria[78], and their activity could prevent the excessive proteolytic activity of intestinal content in IBS. The depletion of bifidobacteria has been noted in both faecal and mucosal samples of patients with IBS[51,79], suggesting an important role for this bacterial genera in IBS. The fermentation of proteins generates numerous health-compromising substances[80]. Among these, hydrogen sulfide is a relevant toxin that impairs epithelial metabolism[81] and can be further converted to tetrathionate, which stimulates the growth of tetrathionate-utilizing pathogens from Gammaproteobacteria[82,83]. The abundance of several Gammaproteobacteria significantly correlates with bowel symptoms in patients with IBS[51,52], and also with the levels of the inflammatory markers interleukin 6 (IL-6) and IL-8 (REF. 51) that are typically increased in IBS[54].

Microbiota and 5-HT. 5-HT is an important metabolite that, among other functions, regulates gastrointestinal motility; disturbed levels of 5-HT seem to be relevant for IBS pathology[84]. As much as 90% of 5-HT is produced in enteroendocrine cells present in the gastrointestinal tract, and it has been recently shown that intestinal bacteria are needed for the stimulation of 5-HT synthesis. Attempts to identify microorganisms that are capable of 5-HT synthesis have shown that, in contrast to *Bacteroides* spp. and altered Schaedler flora (a community of eight bacterial strains), only specific spore-forming commensal bacteria have this feature. The majority of these spore-forming bacteria belong

© 2016 Macmillan Publishers Limited. All rights reserved.

Box 3 | **Structural and functional biomarker candidates in IBS***

Altered motility and stool behaviour
- Altered colonic transit time
- Impaired bile acid transport‡

Mucosal permeability
- Reduced epithelial resistance§
- Reduced expression of ZO1§

Immune imbalance
- Increased numbers of intraepithelial $CD3^+$ lymphocytes§
- Increased mucosal cell density and reactivity§
- Increased nerve mast cell association in the lamia propria region§
- Increased levels of T_H2 cytokines in the blood‖
- *TNFSF15* and *TNF* polymorphisms§,‖
- Increased levels of the pattern recognition receptors TLR2 and TLR4§
- Increased levels of anti-flagellin autoantibodies‖
- Increased levels of histamine and proteases in biopsy supernatants§
- Increased production of IL-1β and TNF by PBMCs‖
- Increased levels of β-defensin 2 antimicrobial peptide‡

Neural plasticity
- Increased nerve fibre density in the epithelium and lamina propria§
- Mostly visceral hypersensitivity, but ≤40% of patients are normosensitive or hyposensitive
- Mucosal biopsy supernatants activate the enteric nervous system independent of stool behaviour or visceral sensitivity§
- Mucosal biopsy supernatants activate sensory fibres and dorsal root ganglion neurons (mostly hypersensitive IBS)
- PBMC supernatants evoke mechanical hypersensitivity involving cytokines and TRPA1

Serotonin metabolism and signalling
- Increased plasma levels of serotonin in IBS-D‖
- Increased enterochromaffin cell density§
- Altered *SERT* expression and polymorphism§
- Serotonin receptor and transporter polymorphisms§

Others
- Increased levels of cysteine and serine proteases‡
- Increased levels of mucosal PARM1§
- Increased levels of BDNF and NGF§
- Increased levels of rectal PYY and somatostatin cell count§
- Altered microbiota diversity and composition‡

BDNF, brain-derived neurotrophic factor; IBS, irritable bowel syndrome; IBS-D, IBS with diarrhoea; IL-1β, interleukin 1β; NGF, nerve growth factor; PARM1, prostate androgen-regulated mucin-like protein 1; PBMC, peripheral blood mononuclear cell; PPY, peptide YY; *SERT*, serotonin reuptake transporter; T_H2, T helper 2; TLR, Toll-like receptor; TNF, tumour necrosis factor; TNFSF15, TNF superfamily member 15; TRPA1, transient receptor potential cation channel subfamily A member 1; ZO1, zonula occludens 1. *Based on data available in REF. 244. ‡In stool. §Intestinal biopsy. ‖In blood.

to the Clostridales class within the Firmicutes phylum. Two recent comprehensive studies[51,85] revealed an increase of the Firmicutes phylum members on the account of the Bacteroidetes members in IBS. Given that the Clostridiales class within the Firmicutes phylum are the most diverse and the most abundant group of the microbiota[70], it is not clear if the observed feature of the IBS microbiota is associated with 5-HT-mediated pathophysiology, but this possible link should certainly be further investigated.

Brain and behaviour

IBS is narrowly defined by recurrent abdominal pain and discomfort associated with altered bowel habits in the absence of an organic origin and/or explanation of symptoms. However, given that IBS is nearly always associated with increased anxiety and patients often show comorbidities with other chronic pain and psychiatric conditions, a more widespread dysregulation of the nervous and immune systems is probably implicated[86].

The brain, the gut and its microbiota and the immune system show reciprocal associations in health and disease. On the one hand, the brain, via the autonomic nervous system and the HPA axis, can influence intestinal motility and fluid secretion[87], intestinal epithelial permeability[25,88,89], immune function[90] and gut microbial composition[91], all of which have been reported to be dysregulated in IBS. On the other hand, several of these peripheral alterations can influence brain structure and function either developmentally or in response to acute perturbations, setting up circular regulatory loops between the gut and the brain[92].

In addition to its role in the bidirectional communications with the gut, the brain plays an essential part in assessing the salience of received or expected interoceptive (sensory) information[93], determining how much of this information is amplified or tuned down, to what degree it is modulated by affect[94] and how much of this interoceptive information from the gut is consciously perceived (visceral sensitivity). One of the best-studied behavioural aspects of IBS-related central processing of gut-related information involves a coping strategy referred to as catastrophizing, a term that refers to a bias towards prediction of a high likelihood of worst outcomes[95]. This measure strongly correlates with the severity of pain symptoms and is a primary treatment target in cognitive–behavioural therapy.

Multimodal brain imaging has made it possible to identify differences in functional (evoked and resting state) and structural (grey matter and white matter tracts) aspects of specific brain networks that provide a neurobiological substrate for previously observed affective and cognitive features of IBS (reviewed in REFS 92,96) (FIG. 6). These networks include the salience, attention, sensorimotor and emotional arousal networks. Profound sex-related differences in these networks have also been identified in both healthy individuals and patients with IBS (reviewed in REF. 96). Cross-sectional correlations of brain networks with several clinical and non-brain biological parameters show a relationship between some of these brain signatures with IBS symptom severity and duration, a history of early adverse life events[93], gut metabolite and microbial composition[97], gene expression profiles in PBMCs[98] and gene polymorphisms[99]. On the basis of these neurobiological findings, a comprehensive IBS pathophysiological model can be formulated (FIG. 6), which includes alterations in the appraisal of and selective attention to interoceptive signals (salience and attentional network), central sensory processing of interoceptive information (sensorimotor network) and engagement of emotional arousal associated with experience and expectation of

© 2016 Macmillan Publishers Limited. All rights reserved.

gut sensations. This disease model not only identifies neurobiological correlates of well-characterized clinical and behavioural features of IBS but also provides a plausible explanation for the common coexistence of IBS with other chronic pain conditions and with increased trait anxiety.

Although these findings have identified disease-relevant brain alterations in patients with IBS, mechanistic and longitudinal studies are required to determine the causality between these factors. For example, are central sensorimotor alterations a consequence of increased signals from the gut, are they the consequence of dorsal horn sensitization by increased descending pain-facilitating signals or are they a genetically determined trait that predisposes individuals to IBS and might be present in asymptomatic relatives[100]? The correlation of gut microbial signatures and PBMC expression profiles with structural alterations in the

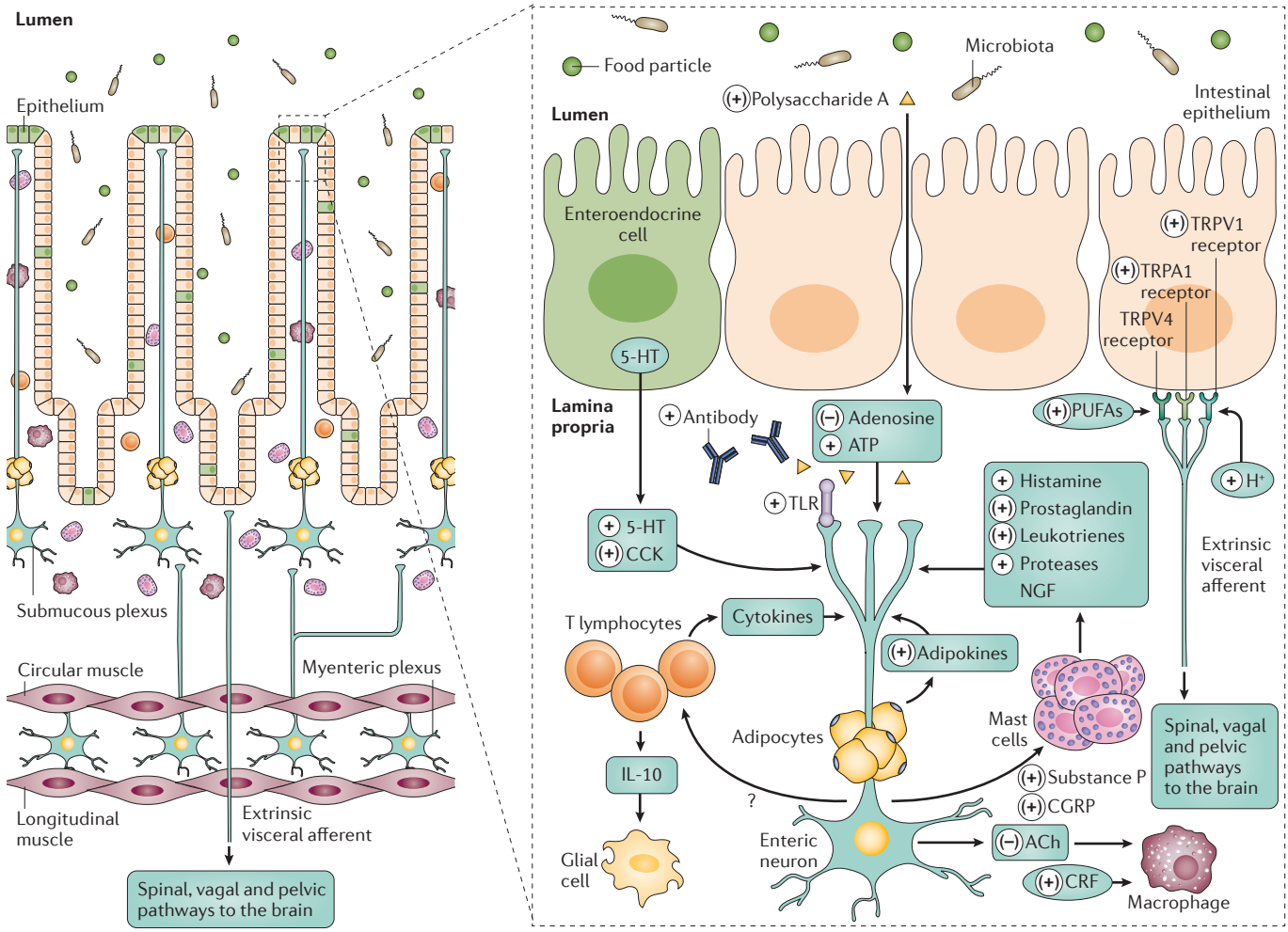


Figure 5 | **Neuroimmune interactions in the gut.** An intimate anatomical and functional association between enteric neurons, terminals from extrinsic nerves and cells of the enteric immune system is the basis for neuroimmune interactions in the gut wall. Functional signalling between nerves and immune cells mostly happens in the epithelial and submucosal layers where there is a high density of immune cells — in particular, T lymphocytes, mast cells and macrophages. The neuroimmune interactions are bidirectional. Enteric neurons, extrinsic nerves and glial cells respond to cytokines and mast cell mediators. Some patients with irritable bowel syndrome (IBS) have circulating autoantibodies against neuronal structures and antibodies that are generated as a response to antigen exposure from the lumen. Neurons can respond directly to antibodies through direct activation of channels or receptors. They also respond to antigens through pathways involving neuronal Toll-like receptor 3 (TLR3), TLR4 and TLR7. Direct signalling between microbiota and the host involves activation of neurons through polysaccharide A. These direct effects of luminal factors are very likely to be outnumbered by signalling between epithelial (in particular, enteroendocrine cells), immune and nerve cells. Neurons also express receptors for adenosine and ATP; both molecules are released in the gut wall under inflammatory or stress conditions. Reciprocally, nerves release factors that affect epithelial or immune cells. The best-documented effect is the activation of mast cells through the release of calcitonin gene-related peptide (CGRP) from extrinsic visceral afferents or enteric neurons. Conversely, acetylcholine (ACh) inhibits the activation of macrophages. Neurogenic inflammation, which is sometimes observed in animal models, is probably caused by the release of CGRP and substance P from extrinsic fibres followed by permeabilization of blood vessels. In addition, adipocytes in the lamina propria nestle against nerve fibres, and release of their pro-inflammatory mediators modulates nerve activity. The plus and minus symbols indicate whether a mediator activates or inhibits its target cell; those in parentheses denote actions established in animal models and those without parentheses are effects demonstrated in humans (human tissue). 5-HT, 5-hydroxytryptamine (also known as serotonin); CCK, cholecystokinin; CRF, corticotropin-releasing factor; IL-10, interleukin 10; NGF, nerve growth factor; PUFA, polyunsaturated fatty acid; TRP, transient receptor potential cation channel.

© 2016 Macmillan Publishers Limited. All rights reserved.

Box 4 | **Dysbiosis in IBS**

Microbiota species increased in IBS

- Enterobacteriaceae
- *Veillonella*
- *Streptococcus**
- *Dorea*
- *Blautia*
- *Roseburia*
- *Ruminococcus*
- *Methanobrevibacter*‡

Microbiota species decreased in IBS

- *Bifidobacterium*
- *Collinsella*
- *Streptococcus*‡
- *Faecalibacterium*
- Christensenellaceae
- Clostridiales
- Uncultured
- *Methanobrevibacter*§

IBS, irritable bowel syndrome. *IBS with diarrhoea. ‡IBS with constipation. §Mixed-type IBS.

sensorimotor network suggests a possible role of these peripheral factors in influencing the brain. Similarly, are the altered salience and attention network alterations a secondary response to the chronically increased perception of visceral signals or are they a primary abnormality that is responsible for the generation of aberrant endogenous pain modulation, as well as emotional and autonomic nervous system responses? Future studies will need to address the question of whether these brain signatures differ between subgroups of patients with IBS, such as male and female patients, patients with a history of early adversity, patients with different durations of symptoms and patients with post-infectious IBS.

Genetic and epigenetic data

The latest genetic and epigenetic findings support current models of IBS pathogenesis that suggest disturbed intestinal barrier function, immune response and neuronal signal transduction[101] (FIG. 6). The data even point towards potential diagnostic biomarkers or therapeutic options (BOX 3). For example, silencing the microRNA-29 (*mir-29*) family or amplifying *mir-199a* expression might have important therapeutic implications for selected patients with IBS and symptoms caused by increased intestinal permeability or hypersensitivity[102,103].

Genetic data. Genetic studies to date range from family and twin studies to candidate gene approaches and, more recently, genome-wide association studies (GWAS). Regardless of enlarged sample sizes, increased statistical power and meta-analyses, genetic variants associated with IBS are still scarce and/or have not been replicated in independent cohorts. A recent paper summarizes all currently available genetic data that have been replicated[101].

Polymorphisms or variants in several genes have been found to be associated with IBS. Genes encoding proteins involved in homeostasis of epithelial barrier function, such as cadherin 1 (*CDH1*) and cell division cycle 42 (*CDC42*), the immune system, such as *IL6, IL10, TNF* and TNF superfamily member 15 (*TNFSF15*; encoding cytokines and neuronal signal transduction) and others (such as neurexophilin 1 (*NXPH1*) and sodium voltage-gated channel α-subunit 5 (*SCN5A*)) have been replicated in several studies[101]. In 2014, a small pilot study reported an association between IBS and a locus on chromosome 10 (containing the protocadherin 15 (*PCDH15*) gene) in a discovery sample from Australia that could not be replicated in additional cohorts from Sweden and the United States[104]. Mutations in the following genes encoding proteins involved in the serotonergic system have also been shown to be associated with IBS: solute carrier family 6 member 4 (*SLC6A4*; also known as 5-HTTLPR or SERT), 5-HT receptor 3A (*HTR3A*), *HTR3E* and *HTR4* (REF. 101). A polymorphism in *SLC6A4* has been found to be associated with altered brain responses, visualized through functional brain imaging following visceral pain stimuli in patients with IBS[105]. Furthermore, a functional polymorphism in *HTR3A* could be associated with altered amygdala responsiveness, anxiety and increased symptom score in IBS[106]. These findings underline the effect of polymorphic serotonergic and other genes in modulating gut-derived brain response in areas that process visceral perception and integrate autonomic control, salience and somatosensory and emotional central networks (FIG. 6).

Variants of genes encoding proteins that are involved in bile acid synthesis regulation (the Klotho-β (*KLB*) gene, the fibroblast growth factor receptor 4 (*FGFR4*) gene and the G protein-coupled bile acid receptor 1 (*GPBAR1*) gene) are associated with accelerated colonic transit in patients with IBS-D[107,108]. These variants also correlate with the colonic transit response to chenodeoxycholic acid (a bile acid used to treat constipation) in IBS-C[109] and to colesevelam (a bile acid sequestrant used to treat diarrhoea) in patients with IBS-D[110,111].

Finally, a locus at 7p22.1 in which the genes KDEL endoplasmic reticulum protein retention receptor 2 (*KDELR2*) and GRID2-interacting protein (*GRID2IP*) localize was significantly associated with IBS risk in the index GWAS (a large twin discovery sample from Sweden) and all replication cohorts in Europe, the United States and Australia[112]. However, the underlying molecular cause for this association finding has not been elucidated.

Epigenetic data. Even less insight into the role of epigenetics in IBS pathology is available compared to the genetic implications. To date, only a few miRNA studies have been performed. These studies reported on the differential expression profiles of miR-29a, miR-29b, miR-103, miR-16, miR-125b and miR-199a in the intestinal mucosa of patients with IBS-D. Upregulation of miR-29a and miR-29b was reported to accompany downregulation of the target genes encoding glutamine synthetase (*GLUL*)[102], claudin 1 (*CLDN1*) and

© 2016 Macmillan Publishers Limited. All rights reserved.

NF-κB-repressing factor (*NKRF*); *CLDN1* and *NKRF* correlated with increased gut permeability[103]. In addition, decreased expression of miR-103, miR-16 and miR-125b correlated with the upregulation of the target genes encoding the tight junction proteins claudin 2 (*CLDN2*) and cingulin (*CGN*)[113]. In turn, a diminished miR-199 level correlated with an upregulation of *TRPV1* and increased visceral sensitivity[114]. Moreover, variants residing in miRNA target regions of the 5-HT receptor genes *HTR3E* and *HTR4B* — namely, c.*76G>A and c.*61T>C — were found to be associated with IBS-D. Both variants were reported to impair miRNA regulation and to lead to disturbed expression regulation of miR-510 and miR-16, respectively[115,116]. One pilot study further indicated increased levels of circulating miR-150 and miR-342-3p in the blood of patients with IBS[117]. Of note, miR-150 has been described to be associated with IBD and pain, whereas miR-342-3p has been predicted to target genes that are relevant for pain signalling, colonic motility and smooth muscle function[118].

Diagnosis, screening and prevention

The diagnosis of IBS relies on the patient fulfilling diagnostic criteria for IBS[119] in conjunction with normal results on a limited number of additional tests and investigations used to rule out other diagnoses with reasonable certainty (FIG. 7). Although a substantial proportion of clinicians[120] prefer a process of thorough exclusion of other diseases, the current recommendation is to base diagnosis on symptoms[119]. There is currently no valid biomarker for IBS[121]. The choice of the tests or investigations deemed necessary to rule out other conditions varies depending on the clinical situation and the symptom profile of the patient. In the majority of cases with a typical clinical history compatible with IBS, only a limited number of laboratory tests are recommended without any need to perform invasive investigations. Screening for IBS risk and for prevention of IBS development is currently not applicable, given the heterogeneity of the disease and the multiplicity of putative pathophysiological mechanisms.

Box 5 | FODMAPs and a low FODMAP diet*

FODMAPs stands for fermentable oligosaccharides (fructans present in wheat, rye, onion and garlic chicory; and galactans present in legumes and beans), disaccharides (lactose present in milk and milk products), monosaccharides (fructose present in artificial sweeteners) and polyols (sugar alcohols present in apples, pears, stone fruit, cauliflower, mushrooms and sweeteners). A low FODMAP diet may include reasonable amounts of:

- Vegetables: bamboo shoots, cucumber, carrot, corn, aubergine (eggplant), lettuce, leafy greens, pumpkin, potato, squash, yam, tomato and courgette (zucchini), among others
- Fruits: banana, cantaloupe, grapes, grapefruit, kiwifruit, kumquat, lemon, lime, mandarin, orange, passion fruit, pawpaw, pineapple, rhubarb and tangerine, among others
- Protein: beef, chicken, canned tuna, egg, egg whites, fish, lamb, pork, shellfish, turkey, cold cuts, nuts and seeds, among others
- Dairy and non-dairy alternatives: lactose-free milk, cream cheese, hard cheeses (cheddar, parmesan and Swiss), mozzarella and sherbet (almond milk, rice milk and rice-milk ice-cream), among others
- Grains: wheat-free grains or wheat-free flours and products made with these (for example, bagels, breads, crackers, noodles, pancakes, pastas, pretzels and waffles), corn flakes, cream of rice, grits, oats, quinoa and rice, among others
- Beverages: water, coffee and tea, and low FODMAP fruit or vegetable juices, among others

*See REFS 171,174.

Diagnostic criteria

As individual symptoms have poor sensitivity and specificity to diagnose IBS, diagnostic criteria incorporating a combination of symptoms have been developed, similar to the DSM system within psychiatry. The first attempt was the so-called Manning criteria, published in 1978 (REF. 122). In this publication, several symptoms were shown to be more common in patients with IBS than in patients with another organic gastrointestinal disease. By combining these symptoms, IBS could be discriminated from other organic gastrointestinal diseases. The experience from the Manning criteria was then used to develop the Rome Foundation criteria, with three different versions over the past 15 years (Rome I, II and III); the latest criteria, the Rome III criteria, was published in 2006 (REFS 119,123,124). The updated Rome IV criteria are expected in May 2016. The sensitivity and specificity of the Rome criteria have been found to be 69–96% and 72–85%, respectively, in different studies, but a problem with these studies is how to define the gold standard for an IBS diagnosis[121].

The common feature in all of these diagnostic criteria is abdominal pain and/or discomfort associated with abnormal bowel habit (diarrhoea (loose and frequent stools), constipation (hard and infrequent stools) or alternating constipation and diarrhoea). All of these criteria require a certain duration and frequency of the symptoms to fulfil the diagnostic criteria for IBS; that is, the symptoms should be chronic and recurring. Thus, the practical clinical use of the diagnostic criteria for IBS involves demonstrating through the clinical history the presence of a combination of these symptoms for ≥3 days per month in the past 3 months, with symptom onset ≥6 months before the diagnosis (Rome III criteria). However, it should be noted that patients with some organic gastrointestinal disease also meet these diagnostic criteria[125] and, as such, the sensitivity and specificity of these criteria is suboptimal to distinguish the different disease entities[125,126].

Clinical features

Besides the symptoms included in the diagnostic criteria, there are other clinical features that support a diagnosis of IBS, even though none of them is mandatory for an IBS diagnosis. One recent study found that variations in stool consistency and frequency or an unpredictable bowel pattern ('irregularly irregular') could be used to discriminate IBS-D adequately from organic gastrointestinal disease[127]. Moreover, abnormal stool frequency (>3 bowel movements per day or <3 bowel movements per week), excessive straining during defaecation, urgency (having to rush to the toilet), feelings of incomplete evacuation and mucus with bowel movements support an IBS diagnosis, but are nonspecific[124].

© 2016 Macmillan Publishers Limited. All rights reserved.

The same is true for postprandial worsening or exacerbation of symptoms, which is common in IBS[128], but is also observed in other gastrointestinal diseases. The presence of other functional gastrointestinal diagnoses (such as functional dyspepsia)[129], as well as reporting numerous functional non-gastrointestinal symptoms and syndromes (such as chronic fatigue, fibromyalgia, uro-gynaecological symptoms, muscle and joint pain and sleep disturbances)[11,130] and psychological comorbidity (such as anxiety and depression)[131], are all common and support an IBS diagnosis.

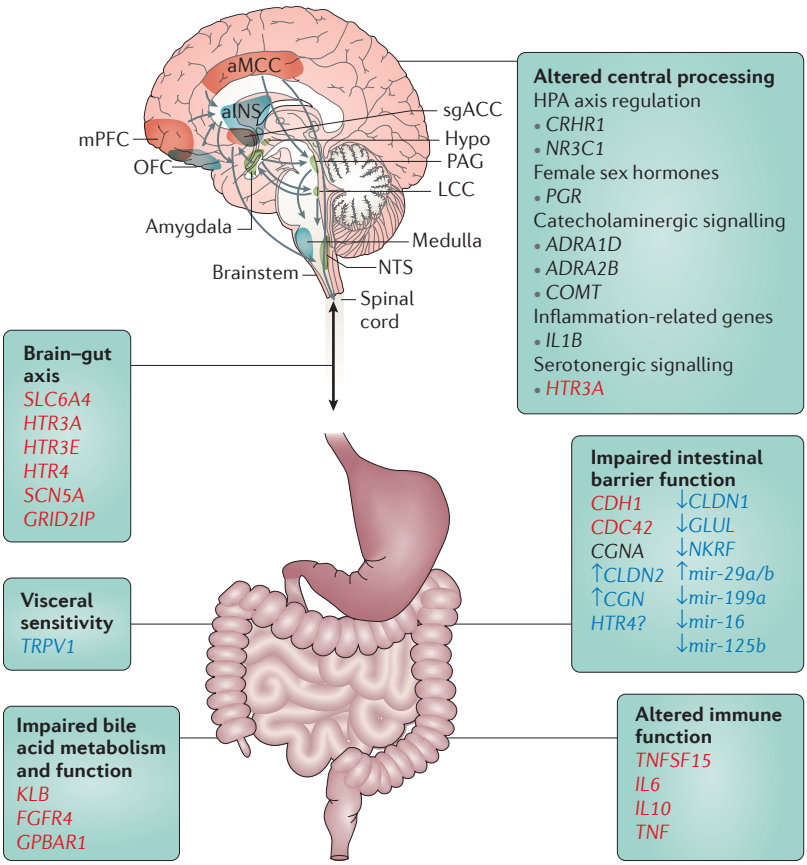


Figure 6 | **Summary of the genetic findings associated with different pathophysiological mechanisms underlying IBS.** Irritable bowel syndrome (IBS)-related pathways that are potential pharmacogenetic targets are marked in red when based on genetic findings and in blue when based on epigenetic findings; those in black are currently not seen as potential pharmacogenetic targets[101]. Various pathways might be affected in specific subgroups of patients with IBS: epithelial barrier (permeability), immune function, impaired bile acid metabolism and function, neuronal processing and signal transduction via spinal afferents from the periphery to the central nervous system in addition to the bidirectional crosstalk via the brain–gut axis, presumably contributing to psychological conditions such as anxiety, depression and somatization. Brain networks that have been associated with structural and functional alteration in IBS are depicted. *ADRA*, adrenoceptor-α; aINS, anterior insula; aMCC, anterior midcingulate cortex; *CDC42*, cell division cycle 42; *CDH1*, cadherin 1; *CGN*, cingulin; *CLDN*, claudin; *COMT*, catechol-*O*-methyltransferase; *CRHR1*, corticotropin-releasing hormone receptor 1; *FGFR4*, fibroblast growth factor receptor 4; *GLUL*, glutamate-ammonia ligase (also known as glutamine synthetase); *GPBAR1*, G protein-coupled bile acid receptor 1; *GRID2IP*, GRID2-interacting protein; HPA, hypothalamus–pituitary–adrenal; *HTR*, 5-hydroxytryptamine receptor; hypo, hypothalamus; *IL*, interleukin; *KLB*, Klotho-β; LCC, locus coeruleus complex; *mir*, microRNA; mPFC, medial prefrontal cortex; *NKRF*, nuclear factor-κB-repressing factor; *NR3C1*, nuclear receptor subfamily 3 group C member 1; NTS, solitary nucleus; OFC, orbitofrontal cortex; PAG, periaqueductal grey; *PGR*, progesterone receptor; *SCN5A*, sodium voltage-gated channel α-subunit 5; sgACC, subgenual anterior cingulate cortex; *SLC6A4*, solute carrier family 6 member 4; *TNF*, tumour necrosis factor; *TNFSF15*, TNF superfamily member 15; *TRPV1*, transient receptor potential cation channel subfamily V member 1.

Physical examination

A physical examination should be part of the evaluation to reassure patients and also to help exclude another organic cause of the symptoms. Admittedly, an abdominal examination, which is part of the routine examination, rarely discloses a specific diagnosis (that is, abdominal tenderness is present in various diseases), but the absence of objective findings on a physical examination has been found to support a diagnosis of IBS[132]. A digital rectal examination is an important part of the physical examination and a useful tool to identify patients with dyssynergic defaecation, which is important to exclude in patients with constipation[133,134] as well as to exclude rectal cancer. Perianal inspection should also be part of the examination to rule out perianal fistulas and other relevant anal pathology.

Laboratory tests

From the existing literature, it is not obvious which laboratory test to recommend in the diagnostic work-up of patients with IBS symptoms. Only serological tests for coeliac disease seem to be more likely to be abnormal in patients with symptoms compatible with IBS than in the general population[135], even though a large multicentre trial failed to confirm this[136]. However, few studies have systematically evaluated the usefulness of laboratory tests in patients with potential IBS. A recent systematic review demonstrated that C-reactive protein (CRP) levels of ≤0.5 mg per dl or faecal calprotectin levels of ≤40 μg per g essentially exclude IBD in patients with IBS symptoms[137]. On the basis of the existing literature, it seems reasonable to perform a complete blood count and CRP measurement, as these are inexpensive and can be used to reassure the health-care provider and the patient. A thyroid profile can be included if the clinical suspicion of thyroid disease is high, a serological test for coeliac disease can be recommended in patients with non-constipated IBS and — if there is suspicion of an inflammatory process — a faecal calprotectin measurement can be added. Stool analyses to detect gastrointestinal infections can be considered if diarrhoea is predominant and difficult to treat, especially in regions where infectious diarrhoea is common[138]. As stated previously, there is currently no valid diagnostic biomarker, even though preliminary data have suggested that certain biomarkers or biomarker assays (BOX 3) for clinical use might prove to be valid following further scientific investigation[139,140].

Alarm features

Alarm features for IBS are symptoms that should raise the clinical concern of another gastrointestinal disease rather than IBS. Whether the use of alarm features (BOX 6) improves the performance of diagnostic criteria for IBS is not totally clear[125,141]. However, from a clinical point of view, it seems reasonable to use these to select patients for further diagnostic testing, even though these may be

© 2016 Macmillan Publishers Limited. All rights reserved.

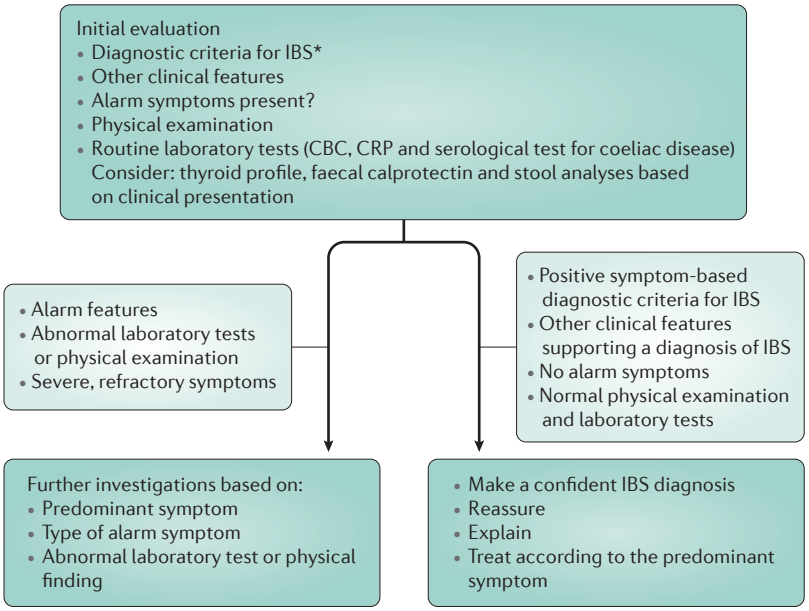


Figure 7 | **A diagnostic algorithm for patients with IBS.** This diagram gives a schematic overview of the sequential approach to irritable bowel syndrome (IBS) diagnosis[144]. CBC, complete blood count; CRP, C-reactive protein. Figure from REF. 144, Nature Publishing Group.

present in a substantial proportion of patients without indicating a serious underlying condition in the gastrointestinal tract[142]. Alarm symptoms can necessitate further investigations to rule out another gastrointestinal disease before an IBS diagnosis can be recommended. Moreover, the predominance of diarrhoea, especially when watery and frequent, should alert the clinician to consider alternative diagnoses[143].

Invasive investigations

In the majority of patients with symptoms compatible with IBS and normal routine laboratory tests but without alarm features[144], no additional invasive investigations are needed and, importantly, performing investigations does not seem to improve patient satisfaction or QOL[145,146].

Colonoscopy should be performed when alarm features prompt an investigation and when there is suspicion of an inflammatory condition in the gastrointestinal tract based on history or laboratory parameters (increased CRP or faecal calprotectin levels)[137], or based on the indications for colorectal cancer screening in countries with population screening programmes[147,148].

When the patient complains of watery diarrhoea as the predominant symptom, a colonoscopy with biopsies should also be considered to rule out microscopic colitis, especially in women >50 years of age[143,149]. Moreover, bile acid-induced diarrhoea has recently been found to be a very important differential diagnosis in patients with IBS symptoms with frequent, loose stools[32,33], and a diagnostic test should be considered (75-homocholic acid taurine (^{75}SeHCAT) test or serum C4 levels)[150]. Unfortunately, these tests are not available in all centres, therefore a therapeutic trial with a bile acid-binding agent is often used as an indirect, but far from perfect, assessment of bile acid-induced diarrhoea.

Carbohydrate malabsorption is another differential diagnosis in patients with IBS-D[151–153], and lactose or fructose hydrogen breath tests can be considered[154,155], but a trial period with dietary exclusion of the suspected carbohydrate for several weeks is often used instead.

If coeliac disease is suspected, based on a positive serological test or the clinical history, an upper gastrointestinal endoscopy with duodenal biopsies should be performed. Small intestinal bacterial overgrowth has been proposed to be common in IBS, but its prevalence and clinical importance is uncertain, therefore routine clinical testing for this cannot be advocated[156,157], especially as valid tests with adequate sensitivity and specificity are lacking.

Management

Only a fraction of patients with IBS-like symptoms (~50%) seek medical care[158]. Most of these patients will initially consult primary care physicians for their symptoms, and the factors that drive this consultation are symptom severity, especially pain, the occurrence of alarm symptoms (BOX 6) and concerns that symptoms might indicate an underlying severe disease — for example, cancer[159]. Therefore, in many cases, gastrointestinal specialist care is needed to exclude diseases that can mimic IBS symptoms — for example, by endoscopy. Once a positive diagnosis of IBS has been established, clinical management can be carried out as well by primary care physicians and at substantially lower costs[160].

Management of IBS involves an integrated approach, including the establishment of an effective patient–provider relationship, education, reassurance, dietary alterations, pharmacotherapy and behavioural and psychological treatment[161]. Owing to the fact that ~50–70% of patients with IBS report additional somatic and psychological symptoms when they are asked[161,162], a stepped-care approach including aspects of cognitive and interpersonal therapy is most appropriate[15]. The initial treatment strategy should be based on predominant symptoms and includes antispasmodics for abdominal pain, antidiarrhoeals for IBS-D and laxatives for IBS-C, whereas nutritional interventions and psychotherapy can be used in all subtypes.

Nutrition

Food ingestion is one of the most commonly reported factors that results in the exacerbation of symptoms among patients with IBS[163,164]. Postprandial symptoms per se and fear of their occurrence (anticipatory anxiety) contribute profoundly to reduced QOL in IBS[128]. Up until recently, food-related symptoms had received scant attention from clinical scientists, leaving patients to find their own way through the plethora of usually non-validated and untested diagnostic tests and dietary regimens, which could result in clinically relevant nutritional deficits[165].

It has become evident that food intolerance (a physiological reaction to food allergens that is not associated with an immune response), and not classical IgE-mediated food allergy (which involves activation of the

© 2016 Macmillan Publishers Limited. All rights reserved.

Box 6 | **Alarm features for IBS**

- Unintended weight loss (>10% in 3 months)
- Presence of blood in the stools not caused by haemorrhoids or anal fissures
- Symptoms that awaken the patient in the night
- Fever in association with the bowel symptoms
- Family history of colorectal cancer, inflammatory bowel disease or coeliac disease
- New onset of irritable bowel syndrome (IBS) symptoms after 50 years of age

immune system), is the major mechanism responsible for symptomatic responses to certain foods[165]. This is not to say that immune responses to food or food components are irrelevant for IBS. For example, one study demonstrated that exposure of the small intestine to certain food antigens led to subtle ultrastructural changes in the duodenal mucosa of patients with IBS, but not in controls[31]. Another study also reported local immune responses to gluten among a group of non-coeliac patients with IBS[166]. Taken together, these observations leave the door open to the possibility that at least some patients with IBS may mount an, as yet to be defined, immunological response to certain dietary components, a response that seems to be confined to the mucosal immune system.

How does one explain food-related symptoms in IBS? Given the primacy of food ingestion as a stimulus to most gastrointestinal functions, postprandial pain and rectal urgency in IBS could simply reflect an exaggeration of a normal physiological phenomenon. Exaggerated motor responses to food and, especially to lipids, have also been demonstrated in the small intestine in IBS[167]. Furthermore, tryptophan, the 5-HT precursor, and related compounds present in some foods could modulate psychological comorbidities and gastrointestinal symptoms in IBS[168]. Food-related symptoms could also be mediated through interactions between our diet, the products of digestion and the gut microbiota. Products of bacterial metabolism, such as deconjugated bile salts, SCFAs and gases, could exert potent effects on colonic physiology and thereby induce symptoms.

Although patients with IBS readily incriminate specific food items as those that are especially likely to precipitate symptoms, only 11–27% of those are correctly identified when confirmed in formal, blinded food challenge studies[169]. The limitations of dietary surveys and the poor reproducibility of reported food intolerances notwithstanding, some food items are reported as being more problematic: wheat, fruit and vegetables[170]. Current enthusiasm for diets low in FODMAPs is consistent with these observations.

Fibre and fibre-based supplements accelerate colon transit, increase stool bulk and facilitate its passage, resulting in an increase in stool frequency. These effects translate into clinically meaningful benefits for people with chronic constipation and IBS-C. Indeed, fibre and products based on synthetic fibre-like substances became a cornerstone in the management of IBS. However, RCTs found that not all patients gained relief and some even complained of exacerbation of their symptoms (including pain, bloating and distension). Recent meta-analyses and systematic reviews have shed some light on this issue by showing that fibres are heterogeneous and the consumption of soluble fibres such as psyllium, calcium polycarbophil and ispaghula bring symptomatic benefits, whereas insoluble fibres, represented by bran, are ineffective in patients with IBS[169].

Interest in the use of low FODMAP diets (BOX 5) in patients with IBS is increasing. RCTs have confirmed some beneficial effects of low FODMAP diets on IBS symptoms[171], but they were not superior to conventional dietary advice when directly compared[172]. There are some limitations; studies to date have been small and, as has been the case with many studies of dietary interventions in IBS, suffer from some methodological limitations[173]. Furthermore, low FODMAP diets are complex, may require supervision by a qualified dietician and involve the elimination of many food items commonly regarded as components of a 'healthy' diet. Some initial investigations suggest that the low FODMAP diet may suppress the growth of bacterial species commonly regarded as important components of healthy microbiota, such as bifidobacteria[174]. Included in the FODMAP category are some molecules, such as lactose, fructose and sorbitol; some patients with IBS may benefit from the removal of one of these substances alone[175]. Predicting responders is difficult, as commonly used challenge tests, such as the lactose or fructose breath hydrogen test, do not seem to be of value[175,176].

The concept of 'non-coeliac gluten sensitivity' has been advanced to explain instances of IBS-type symptoms that develop in individuals who do not satisfy diagnostic criteria for the diagnosis of coeliac disease (that is, positive serology and appropriate changes in small intestinal morphology)[177]. This remains an unsettled and contentious issue with some studies reporting that, when tested in a blinded manner, gluten did induce the usual IBS symptoms in some patients with IBS[178]. Others argue that gluten contributes little to IBS symptomatology, but that fructans (FODMAPs contained in wheat), and not gluten, are the culprits of wheat-related problems. Results of clinical trials assessing the role of gluten exposure in IBS pathology have therefore, not surprisingly, yielded mixed results[179,180]. Although gluten-free diets are currently enjoying considerable popularity among patients with IBS and the population at large in the United States, the rationale for gluten exclusion in IBS has yet to be firmly established.

Patients with IBS commonly consume any one or combinations of a wide variety of dietary supplements ranging from vitamins to 'digestive enzymes', antioxidants and essential oils. Few, if any, of these have been subjected to rigorous study. Prebiotics (non-digestible food ingredients that beneficially affect the host by selectively stimulating the growth and activity of one species or a limited number of species of bacteria in the colon) and probiotics (live microbial food ingredients that alter the microflora and confer health benefit) have also been used for decades in IBS in the absence of supportive data. Prebiotics and probiotics are now subjected to more-rigorous studies, as they might

© 2016 Macmillan Publishers Limited. All rights reserved.

contribute to altered microbiota in IBS[181,182]. Although these studies must be interpreted with care, a recent meta-analysis does suggest efficacy for probiotics (as a category) in IBS[183]. However, high-quality RCTs remain few in number and available data provide scant information to assist the consumer in choosing a particular product to alleviate symptoms[184] or to make a recommendation on prebiotics or synbiotics (a combination of a prebiotic and a probiotic) in IBS[185].

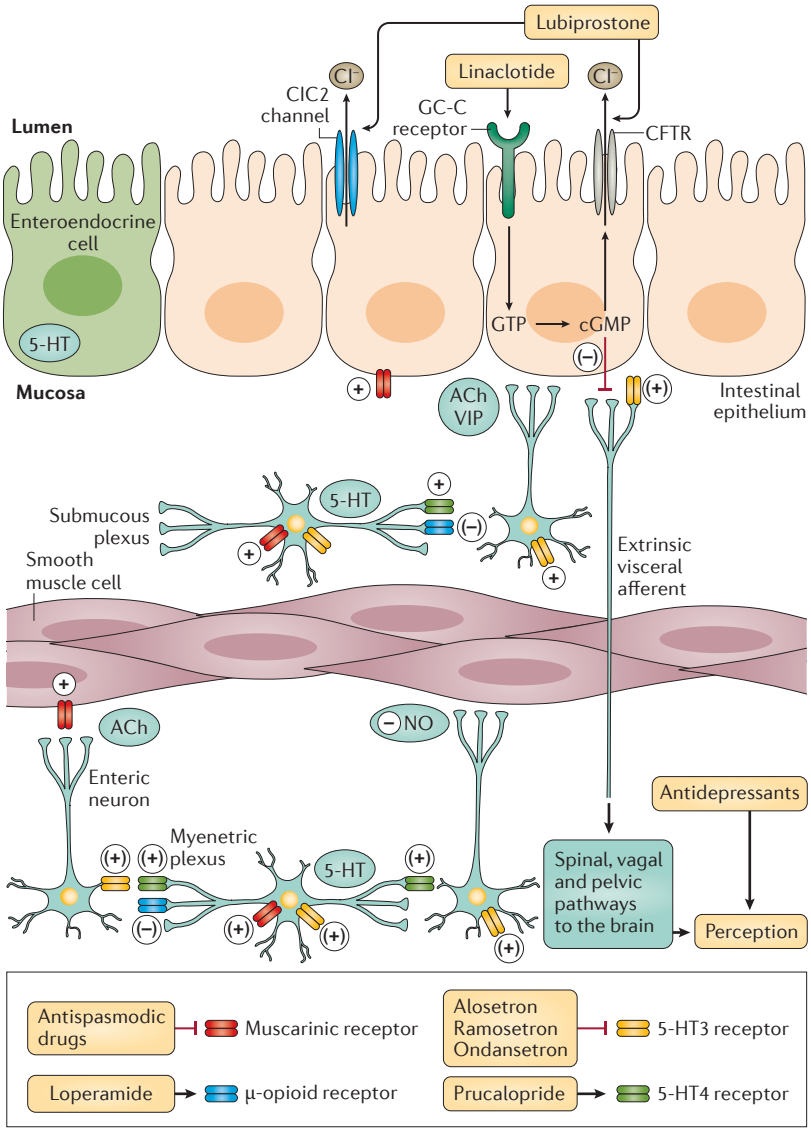


Figure 8 | **Mechanisms of action of different drugs used for the treatment of IBS.** Drugs currently used for the treatment of irritable bowel syndrome (IBS) (orange boxes) target nerve activity, epithelial functions or the contractile state of the smooth muscle layers. Several drugs act by enhancing the activity of chloride channels to increase fluid secretion into the intestinal lumen as a consequence. Other mechanisms of action include modulation of visceral sensitivity at a central or peripheral level. Finally, drugs act to modulate signal transduction at the neuromuscular junction or alter motility by direct myogenic actions. The plus and minus symbols indicate whether a mediator activates or inhibits its target cell; those in parentheses denote actions established in animal models and those without parentheses are effects demonstrated in humans (human tissue). 5-HT, 5-hydroxytryptamine (also known as serotonin); ACh, acetylcholine; CFTR, cystic fibrosis transmembrane conductance regulator; ClC2, chloride channel protein 2; GC-C, guanylyl cyclase C; VIP, vasoactive intestinal peptide.

Drug therapy

Broadly speaking, the current therapeutic armamentarium in IBS aims to alter predominant problematic bowel habits and/or visceral pain. However, an emerging area is manipulation of the gastrointestinal microbiota.

Antispasmodic drugs. Pain in IBS is mediated through central and peripheral mechanisms, and is in part the result of smooth muscle spasms. The mode of action of antispasmodic drugs is probably their ability to antagonize the binding of acetylcholine to the muscarinic receptor at the neuromuscular junction, with smooth muscle relaxation as a consequence[186]. Some studies have demonstrated a beneficial effect of otilonium bromide and hyoscine over placebo, with a number needed to treat (NNT) of four patients[187]. An adverse effect of anti-muscarinic agents is constipation because of their strong inhibition of intraluminal fluid secretion[186]. Accordingly, these drugs are best used in patients without constipation and should be taken 20 minutes before meals to ease postprandial symptoms. Peppermint oil, which also inhibits smooth muscle contraction albeit by calcium channel blockade, is beneficial in reducing IBS symptoms[188]. A recent RCT in patients with IBS-D and IBS-M demonstrated that a novel formulation of peppermint oil, designed to cause a sustained release within the small bowel, was superior to placebo in causing a reduction in total symptoms[189].

Low-dose antidepressants. Antidepressants, such as tricyclic antidepressants (TCAs) or selective serotonin reuptake inhibitors (SSRIs), are recommended by existing guidelines for the treatment of pain in patients who are refractory to antispasmodics and dietary alterations[190]. However, these drugs are not licensed anywhere in the world for the treatment of patients with IBS, and their use is off-label. Given the lack of licensed indication, the rationale for using such drugs should be discussed in detail with patients. The exact analgesic mechanism of action of low-dose antidepressants is incompletely understood but is considered to be both peripheral, via alterations of histaminergic and/or cholinergic transmission within the gastrointestinal tract, and central, via modulation of both ascending visceral sensory afferents and central transmission[191]. SSRIs are generally well tolerated. Adverse effects such as constipation, dry mouth, drowsiness and fatigue are reported with TCAs. TCAs may be particularly effective for treating pain in patients with IBS-D, but are less suitable for patients who have IBS-C.

Laxatives and motility accelerants. In those with constipation, simple laxatives such as senna and docusate are often effective in managing symptoms. However, the use of lactulose is not recommended as it is often poorly tolerated by patients with IBS because of worsening of bloating and discomfort. Linaclotide, a minimally absorbed guanylyl cyclase C agonist peptide (FIG. 8), can be used as second-line therapy after laxatives have failed in patients with IBS-C and symptoms

© 2016 Macmillan Publishers Limited. All rights reserved.

have lasted for >1 year. Linaclotide has a dual action through increasing intraluminal fluid secretion thereby giving its laxation effect but also an analgesic effect via modulation of colonic nociceptors[192], and its effects caused reduced abdominal pain, bloating and bowel symptoms in two well-designed Phase III RCTs[193,194]. Lubiprostone, a minimally absorbed, locally active, bicyclic fatty acid derivative of prostaglandin E1, activates type 2 chloride channels on the enterocyctic apical membrane, thereby stimulating fluid secretion. Lubiprostone has been shown to improve global intestinal symptoms in IBS-C[195]. 5-HT4 receptor agonists (such as prucalopride), which promote gut motility through the activation of the serotoninergic pathways, have been shown to be effective in increasing complete spontaneous bowel movements in patients with chronic constipation[196].

Antidiarrhoeals. The μ-opioid receptor agonist loperamide is frequently used as a first-line agent in IBS-D and improves diarrhoea by inducing peristalsis, which prolongs the gastrointestinal transit time. As loperamide does not cross the blood–brain barrier, central adverse effects are limited. Its main benefit is reducing stool frequency and defaecation urgency, and improving the consistency of the stool[197]. Eluxadoline, a mixed μ-opioid receptor agonist and δ-opioid receptor antagonist, has been evaluated in a Phase III RCT, although safety concerns have been expressed concerning the excess rates of pancreatitis[198].

5-HT3 receptor antagonists, such as alosetron, ramosetron and ondansetron, are effective in the management of IBS-D symptoms. The mechanism of action of 5-HT3 receptor antagonists is complex and incompletely understood, but is considered to proceed through inhibition of the ascending excitatory component of the peristaltic reflex and of the high amplitude propagating contractions within the gastrointestinal tract[199]. However, a central effect of 5-HT3 receptor antagonists on pain cannot be excluded[200]. Safety concerns, with respect to ischaemic colitis, have been confined to alosetron, which subsequently led to restrictions in its prescription[201]. Consequently, other 5-HT3 receptor antagonists have been investigated, with ondansetron[202] and ramosetron demonstrating efficacy in RCTs[203].

Manipulation of the microbiota. Given the burgeoning evidence of the role of the microbiota in IBS, both antibiotics and probiotics have been evaluated. The non-absorbable antibiotic, rifaximin, has been demonstrated to cause a reduction in symptoms, with a NNT of approximately 11 patients, although it is not clear whether repeated courses of treatment are needed[204]. The mechanisms by which rifaximin exerts its positive effects on IBS symptoms are incompletely understood and may include modulation of the gut microbiota, but also direct effects on local micro-inflammation. Rifaximin is approved for use in the United States, but has not yet received regulatory approval in Europe. Probiotics can reduce pain and symptom severity, although recent meta-analyses have highlighted that inconsistencies in study design render definitive recommendations problematic[183,184,205]; again, it is unclear whether probiotics act on IBS symptoms through direct modulation of the microbiota, indirect via the gut immune system or otherwise.

Others. A proportion of patients use herbal supplements either as single herbs or in combination. Four weeks of treatment with iberogast, which is a mixture of nine plant extracts, improved abdominal pain and QOL in a double-blind RCT of 208 patients with all types of IBS[206]. Although the mechanism of action is poorly understood, it is probably multifaceted via acetylcholine, 5-HT and opioid receptors in the gastrointestinal tract[207]. Although herbal remedies represent a promising intervention, further rigorously designed larger RCTs in the subtypes of IBS are needed.

Psychotherapy

The biopsychosocial model of IBS suggests that abdominal symptoms secondarily influence anxiety and depression (bottom-up) and psychosocial factors influence physiological factors, such as motor function, sensory threshold and stress reactivity of the gut (top-down)[208].

Treatment concepts that target these psychosocial factors of patients with IBS should be based on evidence-based models that take the following three components into account: altered peripheral regulation of gut function, altered brain–gut signalling and reducing psychological distress, including general hypervigilance and a general mindset of catastrophizing[209]. Such models might be helpful as a basis of patient education and a target for effective treatments. To further improve treatment programmes, we have to learn more about IBS-specific interactions and the role of stress and visceral sensitivity for clearer evidence on which group of patients might benefit from which treatment approach. In addition, it should be noted that patients with IBS often experience additional functional symptoms, pointing to the complexity of the condition[15].

The effect of IBS symptoms on patients' feelings of shame, fearfulness and embarrassment is well established; patients report being poorly understood by their physicians, as well as by their family members and friends[210]. Patients who experience a positive therapeutic physician–patient relationship have fewer IBS-related follow-up visits[211].

International treatment guidelines for IBS have advocated for a graded treatment approach[212,213]. The National Institute of Health and Care Excellence (NICE) guidelines advise that patients whose symptoms do not respond to pharmacological treatments after 12 months and who develop a continuing symptom profile (refractory IBS) should be considered for referral to cognitive–behavioural therapy (CBT), hypnotherapy (gut-directed hypnosis) or other psychological therapy, such as psychodynamic (interpersonal) therapy and mindfulness-based therapy[190].

© 2016 Macmillan Publishers Limited. All rights reserved.

BOX 7 describes the four major psychological-based therapies for patients with IBS. Several meta-analyses have been performed in the field of psychological and behavioural therapies (including studies in stress reduction and relaxation) that took 45 RCTs into account with a total of 3,325 patients with IBS of all subtypes (TABLE 1). Overall, the NNT for psychological therapies is four patients (95% CI: 3–5) and, therefore, better than the majority of drugs[214]. In a stepped-care approach (beginning with the least intensive or invasive treatment and stepping up or down depending on the needs of the patients), a psychology-based self-aid (educational) approach has been shown recently in a meta-analysis as an effective treatment option for all subtypes of IBS[250]. Compared with control treatments, a medium effect size was demonstrated on decreased symptom severity and a large effect size on increased patient's QOL.

The best evidence is available for CBT. Although CBT is not routinely available in primary care, it can be accessed in some local hospitals and health-care systems. There are medium-to-large significant pooled effect sizes for an improvement of IBS symptoms using CBT with a medium significant pooled effect size for QOL and a small-to-medium pooled effect size for psychological comorbid symptoms. The NNT for CBT is only three patients, with a limited variance between the RCTs. Nevertheless, to date there is no evidence of a superiority of CBT compared with other psychological treatments in IBS.

Box 7 | **Evidence-based psychological treatments**

Cognitive–behavioural therapy

Cognitive–behavioural therapy (CBT) is based on the assumption that irritable bowel syndrome (IBS) symptoms are a response to stressful life events, maladaptive behaviour and an inappropriate attribution of symptoms. CBT aims to modify these behaviours and thoughts through education, which consists of the explanation of IBS symptoms and the CBT model, and by identification of the psychological factors that are interacting with their physical symptoms. On the basis of these findings, patients and therapists work together to identify the potential associations between IBS symptoms and their thoughts, emotions and actions. Finally, behavioural therapy (for example, stress management) is applied[245].

Psychodynamic (interpersonal) therapy

Psychodynamic (interpersonal) therapy (PIT) aims to obtain insights into symptom development as a consequence of interpersonal conflicts or difficulties in relationships with key people. Patients are encouraged to discuss their symptoms in depth, emotional factors are explored and links between symptoms and emotional factors are identified[246].

Gut-directed hypnosis

In gut-directed hypnosis (GDH), as opposed to standard hypnotherapy, suggestions are made on how to control and normalize gastrointestinal function and metaphors are used to bring about improvement. GDH differs from other forms of psychological treatment in which therapy is provided to patients in a conscious state. After information on the effects of hypnosis is given, participants are provided with a compact disk (created by hypnotherapists) for practicing at home on a daily basis[247].

Mindfulness-based therapy

Mindfulness-based therapy (MBT) for IBS has been adapted from the mindfulness-based stress reduction programme. The basic course emphasizes the relevance of mindfulness in coping with IBS-related symptoms and perceptions. With a range of behavioural and cognitive techniques, MBT promotes sensory versus emotional processing of interoceptive signals and counteracts catastrophizing as a maladaptive cognitive coping style[215].

Validation of psychodynamic (interpersonal) therapy, gut-directed hypnosis and mindfulness-based therapy (BOX 7) has only been done in a very limited number of tertiary treatment centres and the generalization of these treatment approaches is limited. Finally, mindfulness-based therapy for IBS shows some promising initial results, particularly in the subgroup of female patients with IBS[215]. Very limited data on multicomponent therapies and on the combination of antidepressants and psychological treatments are available[169]. Overall, there is a lack of reports of adverse effects of psychological and behavioural treatment approaches and treatment resistance in patients with IBS. Psychological therapies have also regularly not distinguished between IBS subtypes and, thus, might have missed differential indications and advantages and disadvantages.

Quality of life

In the field of medicine, general QOL and disease-specific QOL are distinguished. General QOL is a measure of the entire health perception of a person. Representative general QOL can be assessed using the Medical Outcome Study 36-item Short-Form Health Survey (SF-36)[216] or the EuroQOL survey[217]. SF-36 is the most popular instrument that can evaluate physical functioning, physical role, bodily pain, general health perceptions, vitality, social functioning, emotional role and mental health[216]. Disease-specific QOL is a measure of life disturbance that is specifically caused by the disease[218,219].

QOL in patients with IBS is greatly disturbed. Patients with IBS showed impaired general QOL with lower values on all SF-36 subscales except physical functioning than healthy controls in one study[220], whereas lower values on the SF-36 subscales in patients with IBS (except physical functioning, physical role and emotional role) than in healthy controls were observed in another study[221]. All subscales of SF-36, except the physical functioning and physical role domains, were lower in patients with IBS than in healthy controls regardless of culture[222]. The degree of disturbance of general QOL in patients with IBS has been shown to be worse in several subcategories than in those with gastroesophageal reflux disease, diabetes mellitus or severe chronic kidney disease[220]. Finally, a study has shown that patients with IBS had more disturbed general QOL in physical role, bodily pain, general health perceptions and social functioning than non-consulters with IBS (individuals who do not seek treatment)[221].

QOL seems to be the same among IBS subtypes. However, disease-specific QOL, as measured with the IBS-QOL in patients with IBS-D or IBS-M, was worse than in patients with IBS-C in one study[222]. In this study, increased food avoidance in patients with IBS-D and IBS-M may have been responsible for the lower QOL[222], but there are controversial reports[218].

In severe IBS, both gastrointestinal symptoms and psychiatric comorbidity independently contribute to disturbed QOL[223] (FIG. 9). Another study revealed that the QOL of patients with IBS was more influenced by the extraintestinal symptoms — such as tiring easily, low in energy, the feeling that there is something

© 2016 Macmillan Publishers Limited. All rights reserved.

seriously wrong with their body, feeling tense, feeling nervous, feeling hopeless, difficulty sleeping and low sexual interest — than by gastrointestinal symptoms[224]. The psychological and psychosocial dimensions of food ingestion might also have a role. Eating with family and friends is probably the most common form of social interaction worldwide. An inability to participate in such a fundamental component of social intercourse because of a fear of pain, urgency, diarrhoea or distension occurring during or immediately after a meal can be devastating and can result in social isolation[210].

Systematic reviews have clarified that improvement of IBS-related pain by treatment results in better QOL in patients with IBS[225]. The disease-specific IBS-QOL and IBS-QOL questionnaires can measure the efficacy of treatment, especially long-term therapies[226]. Although the SF-36 can also detect the efficacy of long-term treatment (>1 year), it is less sensitive than the IBS-QOL. Both measures struggle to detect drug or psychotherapy efficacy in the short term (<1 month)[226,227], but IBS-QOL is sensitive enough to detect efficacy for mid-term (3 months) treatment[203]. A therapeutic gain of ≥14 points in IBS-QOL denotes a clinically meaningful change. Even if primary end points based on cardinal symptoms of IBS are similar between treatments, a treatment resulting in better QOL may be preferred by patients over another treatment that does not improve QOL.

Outlook

The field of research into IBS has expanded considerably over the past decade with many new studies, in part driven by the development of new therapeutic agents. This trajectory seems likely to continue as patients with IBS account for a substantial proportion of all gastrointestinal consultations, and many questions in the field remain unanswered (BOX 8).

Patient stratification and biomarkers

Many classes of drugs have been evaluated by RCTs in IBS and these have often produced disappointingly small differences from placebo[187,214,228]. These small differences conceal the fact that some patients benefit from the drugs.

Table 1 | **Evidence-based psychological treatments for IBS**

Psychological treatment approach*	*n* of studies (*n* of participants)	Main findings	Comments
CBT[248]	18 RCTs (1,380)	• Symptom score: medium-to-large significant pooled effect size‡ (0.67) • QOL: medium significant pooled effect size (0.48) • Psychological distress (depression and anxiety): small-to-medium pooled effect size (0.21) • NNT for CBT was 3 (95% CI: 2–6)	• CBT was superior to waiting lists, basic support or medical treatment alone at the end of treatment but not superior to other psychological treatments
PIT[249]	2 RCTs (273)	Both studies compared PIT with 'supportive listening' applied by the same therapist. Compared with controls: • PIT significantly improved symptoms • PIT showed a large cost-effectiveness • PIT was widely acceptable • PIT significantly improved QOL • PIT significantly reduced costs • The calculated OR for benefit was 2.92 (95% CI: 1.76–4.83) • NNT for dynamic psychotherapy was 3.5 (95% CI: 2–25)	• PIT is less well standardized in terms of its performance (that is, duration, setting and phases)
GDH[247]	7 RCTs (452)	• 6 of 7 RCTs reported a significant reduction (all $P < 0.05$) in overall gastrointestinal symptoms compared with supportive therapy only • Response rates ranged between 24% and 73% • Efficacy was maintained long term in four of five studies • NNT was 4 (95% CI: 3–8)	• Very few professionals are trained for the specific implementation of GDH and therefore their services can be difficult to access • The mechanisms by which GDH exerts its effect are poorly understood
MBT[215]	2 RCTs (79)	• Women showed greater reductions of symptoms compared with a control group immediately after training (26.4% versus 6.2%; $P = 0.006$) and at 3 months follow-up (38.2% compared with 11.8%; $P = 0.001$) • Changes in QOL, distress and anxiety were not different between groups immediately after treatment • Significantly greater improvement in the MBT group than in the control group evident at 3 months follow-up • The beneficial effects persisted for ≥3 months	• In another RCT, the IBS symptom severity in the mindfulness-based stress reduction group was not retained at 6 months follow-up
Relaxation[214§]	6 RCTs (255)	• Overall, no benefit of relaxation training or therapy in IBS was detected in the RCTs	• The field of studies on relaxation techniques is diverse
GSHs[250]	10 RCTs (886)	• Compared with control conditions, a moderate effect size on symptom severity (0.72) and a large effect size on the increase of patients' QOL (0.84) was found	• GSHs might be an easily accessible and a cost-effective treatment alternative. However, there is a wide heterogeneity and variance in its performance

The NNT data are based on Ford *et al.*[214]. CBT, cognitive–behavioural therapy; GDH, gut-directed hypnosis; GSH, guided self-help intervention; IBS, irritable bowel syndrome; MBT, mindfulness-based therapy; NNT, number needed to treat; OR, odds ratio; PIT, psychodynamic (interpersonal) therapy; QOL, quality of life; RCT, randomized controlled trial. *See REF. 245. ‡Effect size (for example, Cohen's d): effect sizes of 0.2–0.5 are regarded as small, between 0.5 and 0.8 as moderate and >0.8 as large. §Methods and techniques applied are progressive muscle relaxation, biofeedback and transcendental or yoga meditations.

© 2016 Macmillan Publishers Limited. All rights reserved.

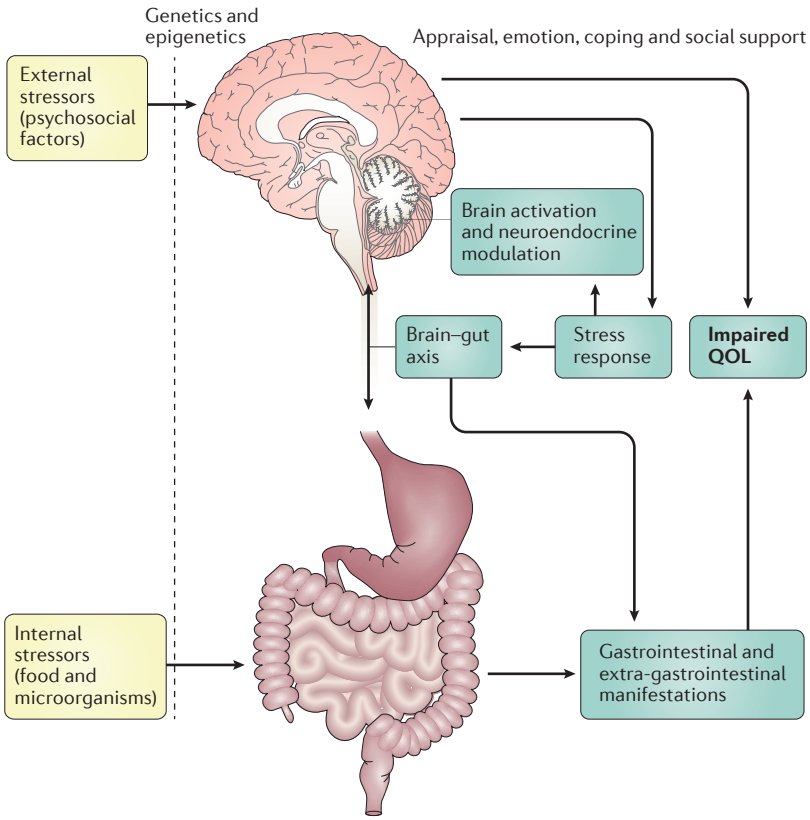


Figure 9 | **Concept of multifactorial quality-of-life effects in IBS.** The genome and epigenome partially determine ('filter') the response of an individual to external stressors (psychosocial factors) and internal stressors (ingested food or microbiota). These, together with social support, appraisal, emotion and coping behaviours against stressors, determine the stress response affecting the brain–gut interactions. This response might involve regional brain activation, changes in autonomic and neuroendocrine function, which might lead to many of the clinical manifestations observed in irritable bowel syndrome (IBS), including visceral hypersensitivity, alteration in gastrointestinal motility, increased mucosal permeability and low-grade inflammation. These gastrointestinal symptoms and other extra-intestinal manifestations (such as multiple somatic symptoms and psychiatric comorbidities) impair the quality of life (QOL) of patients with IBS.

Proper stratification of patients by relevant underlying disease mechanism has been an issue, therefore many trials use unselected patients with IBS, independent of the underlying disease mechanisms and clinical presentations. The use of 5-HT receptor-modulating drugs has taught the research community that restricting 5-HT3 receptor antagonists to patients without constipation improved their effectiveness with significant differences from placebo[229,230], owing to the fact that 5-HT3 receptor antagonists slow transit and aggravate constipation. However, RCTs rarely measure transit as a requirement for trial entry, which depends on symptoms recorded in daily symptom diaries. The use of more-objective biomarkers to select patients for RCTs would be expected to improve the effect size and reduce the number needed to test to show a significant difference from placebo.

The lack of reproducible, widely available biomarkers that reflect the targets of 'older' drugs has been a considerable limitation. Antispasmodics are a good example of such drugs that have fallen out of favour because we cannot reliably identify those with excessive motor activity who might be expected to respond. Future novel non-invasive motility assessments, such as MRI[231], capsule endoscopy[232] and the pressure-sensitive, temperature-sensitive and pH-sensitive SmartPill (Given Imaging Ltd, Yoqneam, Israel)[233] (which can measure intestinal contractions), hold the possibility of identifying such patients in the future.

Although individual genetic markers seem likely to be associated with only quite modest increases in risk for IBS, they might be important predictors of drug sensitivity in particular pathways. 5-HT3 receptor antagonists are good examples of drugs with a wide range of sensitivities such that effective doses for one patient can produce unacceptable constipation in another. This finding may be due to a combination of important functional polymorphisms in genes involved in 5-HT synthesis (tryptophan hydroxylase 1 (*TPH1*)), those involved in 5-HT reuptake via the 5-HT transporter (*SLC6A4*) and polymorphisms in the 5-HT3 receptor genes (which alter sensitivity). Several small studies have suggested significant differences in responder status to one 5-HT3 receptor antagonist, ramosetron, according to polymorphisms in *TPH1* (REF. 234) and to another 5-HT3 receptor antagonist, alosetron, according to polymorphisms in *SLC6A4* (REF. 235). However, these studies are underpowered and have not yet been reproduced[202]. By analogy with other complex disorders[236], the effect of any one individual polymorphism may be limited but combining polymorphisms that predict low 5-HT production with rapid uptake and low receptor sensitivity would be expected to be associated with higher odds ratios for success of 5-HT manipulation. Future studies should be powered to examine this notion such that the dose can be tailored to individual patients. Similarly, polymorphisms in the *FGF19*–*FGFR4* pathway, which controls bile acid synthesis[107,108], influence colonic transit and should be explored to see if different combinations alter sensitivity to bile acid sequestrants or bile acid transporter inhibitors.

Mode of action of food intolerances

Dietary restrictions such as low FODMAP diets (BOX 5) are another example in which implementation of an effective treatment is hampered by lack of biomarkers to predict response or reliably identify the key component (or components) of food that are responsible for symptoms. Although poorly absorbed fermentable carbohydrates can undoubtedly cause symptoms in some patients, visceral sensitivity is the key to why some individuals experience symptoms and some do not[237], at least in the case of lactose malabsorption. However, no trial of lactose exclusion in IBS has used measures of sensitivity to stratify patients. While rectal barostat tests to assess visceral sensitivity are difficult, although not impossible to standardize across centres, alternatives might be to use simple cutaneous pressure or thermal stimulation[238]. More remotely, somatization questionnaires concerning non-gastrointestinal symptoms such as headache, backache, dyspnoea and palpitations have been shown to correlate, albeit weakly, with rectal distension pressure thresholds for pain[239].

© 2016 Macmillan Publishers Limited. All rights reserved.

Box 8 | **Key questions to be addressed in future research**

Can we develop clinically applicable biomarkers to stratify patients to disease mechanisms, thereby reducing the number of patients needed to evaluate new therapeutic agents? Possible factors that should be taken into account are:

- Transit time
- Evidence of bile acid excess
- Immune activation
- Biopsy supernatant mediators that activate enteric neurons
- Mucosal serotonin availability

Can we assess the role of genetic markers in irritable bowel syndrome? Possible factors that should be taken into account are:

- Gene and environment interactions
- Biomarker discovery — for example, by genome-wide association studies
- Pharmacogenetics

Can we identify the mode of action of food intolerances to allow rational designs of diets? Possible tests are:

- Nutrient challenge meals
- MRI studies of intestinal volumes and gas or water content of the stool

Can we characterize the functional effects of changes in microbiota to improve efficacy of manipulation of the microbiota as a novel therapy? Possible studies are:

- Randomized controlled trials of fermentable oligosaccharides, disaccharides, monosaccharides and polyols (FODMAP) intervention with assessment of changes in microbiota
- Effect of placebo-controlled diets on faecal or serum bacterial metabolites

The physical form of food is another key variable whose importance is yet to be defined. Many of the dietary components implicated in IBS symptoms are actually consumed as solids and hence delivered into the duodenum more slowly after trituration by antral contractions. The rapid entry of osmotically active poorly absorbed substrates — mainly in liquid form — such as lactose in a patient with lactose malabsorption[240,241] or mannitol in healthy volunteers[241] result in a rapid influx of water into the small intestine, which probably stimulates transit and rapid delivery into the colon. This leads to the virtually instantaneous generation of gas[242], mainly hydrogen, given that the microbiota are unable to fully metabolize the sudden excess of substrate. Furthermore, distension of the ascending colon generates propulsive colonic motility, which a sensitized individual may experience as cramps; a slower delivery in a solid matrix may be better tolerated. Future studies should define how the physical form of FODMAPs alters their tolerability, which would allow a less restrictive diet that may be easier to follow and, hence, more widely adopted than at present.

Functional effect of changes in microbiota

Many studies have found profound differences in the microbiota of selected patients with IBS, but the agreement on the involved species between studies is poor[57]. Given the very large number of different species that have overlapping metabolic capabilities and functional effects, focusing on function may be more helpful than just identifying the species present.

Analysis of urine and stool metabolites, including bile acids and endogenous tryptase, may provide simpler biomarkers of function that could predict responsiveness to microbiota manipulation. Thus, low levels of butyrate, a SCFA, might encourage the provision of prebiotics that favour butyrate-producing bacteria, such as *Eubacterium rectale* and *Roseburia cecicola*. Future studies should also take into account the important role of transit time and its variability. The challenge of rapid transit favours organisms with either enhanced growth capacity or those that adhere to the mucosa to deal with rapid flow within the colon[243], although, these results need to be replicated and studied in more detail to enable dissection of the extent to which differences in microbiota are the cause or the effect of rapid transit. Better insight might also enable the tailoring of diet to the existing microbiota in a patient, based on their metabolic capabilities and response to a substrate provided in the diet.

1. Lovell, R. M. & Ford, A. C. Global prevalence of and risk factors for irritable bowel syndrome: a meta-analysis. *Clin. Gastroenterol. Hepatol.* **10**, 712–721.e4 (2012).
This meta-analysis covers epidemiological population-based data across 90 studies in 33 countries worldwide and reports not only the pooled prevalences but also several risk factors common to all studies.
2. Gwee, K. A. Irritable bowel syndrome in developing countries — a disorder of civilization or colonization? *Neurogastroenterol. Motil.* **17**, 317–324 (2005).
3. Olafsdottir, L. B., Gudjonsson, H., Jonsdottir, H. H. & Thjodleifsson, B. Stability of the irritable bowel syndrome and subgroups as measured by three diagnostic criteria — a 10-year follow-up study. *Aliment. Pharmacol. Ther.* **32**, 670–680 (2010).
4. Hungin, A. P., Whorwell, P. J., Tack, J. & Mearin, F. The prevalence, patterns and impact of irritable bowel syndrome: an international survey of 40,000 subjects. *Aliment. Pharmacol. Ther.* **17**, 643–650 (2003).
5. Halder, S. L. *et al.* Natural history of functional gastrointestinal disorders: a 12-year longitudinal population-based study. *Gastroenterology* **133**, 799–807 (2007).
6. Ford, A. C., Forman, D., Bailey, A. G., Axon, A. T. & Moayyedi, P. Irritable bowel syndrome: a 10-yr natural history of symptoms and factors that influence consultation behavior. *Am. J. Gastroenterol.* **103**, 1229–1239 (2008).
7. Yarandi, S. S., Nasseri-Moghaddam, S., Mostajabi, P. & Malekzadeh, R. Overlapping gastroesophageal reflux disease and irritable bowel syndrome: increased dysfunctional symptoms. *World J. Gastroenterol.* **16**, 1232–1238 (2010).
8. Ford, A. C. *et al.* Characteristics of functional bowel disorder patients: a cross-sectional survey using the Rome III criteria. *Aliment. Pharmacol. Ther.* **39**, 312–321 (2014).
9. Long, M. D. & Drossman, D. A. Inflammatory bowel disease, irritable bowel syndrome, or what?: A challenge to the functional-organic dichotomy. *Am. J. Gastroenterol.* **105**, 1796–1798 (2010).
10. Keohane, J. *et al.* Irritable bowel syndrome-type symptoms in patients with inflammatory bowel disease: a real association or reflection of occult inflammation? *Am. J. Gastroenterol.* **105**, 1789–1794 (2010).
11. Whitehead, W. E. *et al.* Comorbidity in irritable bowel syndrome. *Am. J. Gastroenterol.* **102**, 2767–2776 (2007).
12. Matheis, A., Martens, U., Kruse, J. & Enck, P. Irritable bowel syndrome and chronic pelvic pain: a singular or two different clinical syndrome? *World J. Gastroenterol.* **13**, 3446–3455 (2007).
13. Janssens, K. A., Zijlema, W. L., Joustra, M. L. & Rosmalen, J. G. Mood and anxiety disorders in chronic fatigue syndrome, fibromyalgia, and irritable bowel syndrome: results from the LifeLines Cohort study. *Psychosom. Med.* **77**, 449–457 (2015).
14. Barsky, A. J. Assessing the new DSM-5 diagnosis of somatic symptom disorder. *Psychosom. Med.* **78**, 2–4 (2016).
15. Henningsen, P., Zipfel, S. & Herzog, W. Management of functional somatic syndromes. *Lancet* **369**, 946–955 (2007).
16. Lovell, R. M. & Ford, A. C. Effect of gender on prevalence of irritable bowel syndrome in the community: systematic review and meta-analysis. *Am. J. Gastroenterol.* **107**, 991–1000 (2012).
17. Goodwin, L., White, P. D., Hotopf, M., Stansfeld, S. A. & Clark, C. Life course study of the etiology of self-reported irritable bowel syndrome in the 1958 British birth cohort. *Psychosom. Med.* **75**, 202–210 (2013).
18. Saito, Y. A. *et al.* Familial aggregation of irritable bowel syndrome: a family case–control study. *Am. J. Gastroenterol.* **105**, 833–841 (2010).
19. Bengtson, M. B., Aamodt, G., Vatn, M. H. & Harris, J. R. Co-occurrence of IBS and symptoms of anxiety or depression, among Norwegian twins, is influenced by both heredity and intrauterine growth. *BMC Gastroenterol.* **15**, 9 (2015).
20. van Tilburg, M. A. *et al.* Psychosocial mechanisms for the transmission of somatic symptoms from parents to children. *World J. Gastroenterol.* **21**, 5532–5541 (2015).
21. Koloski, N. A. *et al.* Identification of early environmental risk factors for irritable bowel syndrome and dyspepsia. *Neurogastroenterol. Motil.* **27**, 1317–1325 (2015).

© 2016 Macmillan Publishers Limited. All rights reserved.

22. Schwille-Kiuntke, J., Frick, J. S., Zanger, P. & Enck, P. Post-infectious irritable bowel syndrome — a review of the literature. *Z. Gastroenterol.* **49**, 997–1003 (2011).
23. Schwille-Kiuntke, J., Mazurak, N. & Enck, P. Systematic review with meta-analysis: post-infectious irritable bowel syndrome after travellers' diarrhoea. *Aliment. Pharmacol. Ther.* **41**, 1029–1037 (2015).
24. Thabane, M., Kottachchi, D. T. & Marshall, J. K. Systematic review and meta-analysis: the incidence and prognosis of post-infectious irritable bowel syndrome. *Aliment. Pharmacol. Ther.* **26**, 535–544 (2007).
This is so far the largest meta-analysis of 18 reports of IBS occurrence following an acute gastrointestinal viral or bacterial infection (called post-infectious IBS) and its potential risk factors and predictors.
25. Bischoff, S. C. *et al.* Intestinal permeability — a new target for disease prevention and therapy. *BMC Gastroenterol.* **14**, 189 (2014).
26. Martinez, C. *et al.* Diarrhoea-predominant irritable bowel syndrome: an organic disorder with structural abnormalities in the jejunal epithelial barrier. *Gut* **62**, 1160–1168 (2013).
27. Piche, T. *et al.* Impaired intestinal barrier integrity in the colon of patients with irritable bowel syndrome: involvement of soluble mediators. *Gut* **58**, 196–201 (2009).
28. Bertiaux-Vandaele, N. *et al.* The expression and the cellular distribution of the tight junction proteins are altered in irritable bowel syndrome patients with differences according to the disease subtype. *Am. J. Gastroenterol.* **106**, 2165–2173 (2011).
29. Coeffier, M. *et al.* Increased proteasome-mediated degradation of occludin in irritable bowel syndrome. *Am. J. Gastroenterol.* **105**, 1181–1188 (2010).
30. Buhner, S. *et al.* Activation of human enteric neurons by supernatants of colonic biopsy specimens from patients with irritable bowel syndrome. *Gastroenterology* **137**, 1425–1434 (2009).
This paper reports the relevance of a novel technique to identify biomarkers of micro-inflammation in IBS by using supernatants from biopsies of patients with IBS applied *in vitro* to isolated neurons of the ENS of animals.
31. Fritscher-Ravens, A. *et al.* Confocal endomicroscopy shows food-associated changes in the intestinal mucosa of patients with irritable bowel syndrome. *Gastroenterology* **147**, 1012–1020.e4 (2014).
32. Slattery, S. A., Niaz, O., Aziz, Q., Ford, A. C. & Farmer, A. D. Systematic review with meta-analysis: the prevalence of bile acid malabsorption in the irritable bowel syndrome with diarrhoea. *Aliment. Pharmacol. Ther.* **42**, 3–11 (2015).
33. Bajor, A., Tornblom, H., Rudling, M., Ung, K. A. & Simren, M. Increased colonic bile acid exposure: a relevant factor for symptoms and treatment in IBS. *Gut* **64**, 84–92 (2015).
34. Valentin, N. *et al.* Biomarkers for bile acid diarrhoea in functional bowel disorder with diarrhoea: a systematic review and meta-analysis. *Gut* http://dx.doi.org/10.1136/gutjnl-2015-309889 (2015).
35. Barbara, G., Cremon, C. & Stanghellini, V. Inflammatory bowel disease and irritable bowel syndrome: similarities and differences. *Curr. Opin. Gastroenterol.* **30**, 352–358 (2014).
36. Barbara, G. *et al.* The immune system in irritable bowel syndrome. *J. Neurogastroenterol. Motil.* **17**, 349–359 (2011).
37. Barbara, G. *et al.* Randomised controlled trial of mesalazine in IBS. *Gut* **65**, 82–90 (2016).
38. Lam, C. *et al.* A mechanistic multicentre, parallel group, randomised placebo-controlled trial of mesalazine for the treatment of IBS with diarrhoea (IBS-D). *Gut* **65**, 91–99 (2016).
39. Barbara, G. *et al.* Activated mast cells in proximity to colonic nerves correlate with abdominal pain in irritable bowel syndrome. *Gastroenterology* **126**, 693–702 (2004).
40. Cenac, N. *et al.* Quantification and potential functions of endogenous agonists of transient receptor potential channels in patients with irritable bowel syndrome. *Gastroenterology* **149**, 433–444.e7 (2015).
41. Nasser, Y., Boeckxstaens, G. E., Wouters, M. M., Schemann, M. & Vanner, S. Using human intestinal biopsies to study the pathogenesis of irritable bowel syndrome. *Neurogastroenterol. Motil.* **26**, 455–469 (2014).
42. Barbara, G. *et al.* Mast cell-dependent excitation of visceral-nociceptive sensory neurons in irritable bowel syndrome. *Gastroenterology* **132**, 26–37 (2007).
43. Cenac, N. *et al.* Role for protease activity in visceral pain in irritable bowel syndrome. *J. Clin. Invest.* **117**, 636–647 (2007).
44. Annahazi, A. *et al.* Fecal proteases from diarrheic-IBS and ulcerative colitis patients exert opposite effect on visceral sensitivity in mice. *Pain* **144**, 209–217 (2009).
45. Annahazi, A. *et al.* Luminal cysteine-proteases degrade colonic tight junction structure and are responsible for abdominal pain in constipation-predominant IBS. *Am. J. Gastroenterol.* **108**, 1322–1331 (2013).
46. Buhner, S. *et al.* Neuronal activation by mucosal biopsy supernatants from irritable bowel syndrome patients is linked to visceral sensitivity. *Exp. Physiol.* **99**, 1299–1311 (2014).
47. Valdez-Morales, E. E. *et al.* Sensitization of peripheral sensory nerves by mediators from colonic biopsies of diarrhea-predominant irritable bowel syndrome patients: a role for PAR2. *Am. J. Gastroenterol.* **108**, 1634–1643 (2013).
48. Hughes, P. A. *et al.* Sensory neuro-immune interactions differ between irritable bowel syndrome subtypes. *Gut* **62**, 1456–1465 (2013).
49. Dothel, G. *et al.* Nerve fiber outgrowth is increased in the intestinal mucosa of patients with irritable bowel syndrome. *Gastroenterology* **148**, 1002–1011.e4 (2015).
50. Turnbaugh, P. J. *et al.* The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature* **449**, 804–810 (2007).
51. Rajilić-Stojanović, M. *et al.* Global and deep molecular analysis of microbiota signatures in fecal samples from patients with irritable bowel syndrome. *Gastroenterology* **141**, 1792–1801 (2011).
52. Jalanka-Tuovinen, J. *et al.* Faecal microbiota composition and host-microbe cross-talk following gastroenteritis and in postinfectious irritable bowel syndrome. *Gut* **63**, 1737–1745 (2014).
53. Saulnier, D. M. *et al.* Gastrointestinal microbiome signatures of pediatric patients with irritable bowel syndrome. *Gastroenterology* **141**, 1782–1791 (2011).
54. Scully, P. *et al.* Plasma cytokine profiles in females with irritable bowel syndrome and extra-intestinal co-morbidity. *Am. J. Gastroenterol.* **105**, 2235–2243 (2010).
55. Malinen, E. *et al.* Association of symptoms with gastrointestinal microbiota in irritable bowel syndrome. *World J. Gastroenterol.* **16**, 4532–4540 (2010).
56. Rajilić-Stojanović, M. *et al.* Intestinal microbiota and diet in IBS: causes, consequences, or epiphenomena? *Am. J. Gastroenterol.* **110**, 278–287 (2015).
This narrative review summarizes state-of-the-art knowledge on the role of the gut microbiota and its influence through dietary, probiotic and pharmacological manipulation.
57. Simrén, M. *et al.* Intestinal microbiota in functional bowel disorders: a Rome foundation report. *Gut* **62**, 159–176 (2013).
58. Crouzet, L. *et al.* The hypersensitivity to colonic distension of IBS patients can be transferred to rats through their fecal microbiota. *Neurogastroenterol. Motil.* **25**, e272–e282 (2013).
59. De Palma, G. *et al.* The adoptive transfer of anxiety and gut dysfunction from IBS patients to axenic mice through microbiota transplantation. *Gastroenterology* **146**, S-845 (2014).
60. Rajilić-Stojanović, M., Heilig, H. G., Tims, S., Zoetendal, E. G. & de Vos, W. M. Long-term monitoring of the human intestinal microbiota composition. *Environ. Microbiol.* **15**, 1146–1159 (2013).
61. David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559–563 (2014).
62. Cummings, J. H. & Macfarlane, G. T. Role of intestinal bacteria in nutrient metabolism. *Clin. Nutr.* **16**, 3–11 (1997).
63. Hamer, H. M. *et al.* Review article: the role of butyrate on colonic function. *Aliment. Pharmacol. Ther.* **27**, 104–119 (2008).
64. Shepherd, S. J., Parker, F. C., Muir, J. G. & Gibson, P. R. Dietary triggers of abdominal symptoms in patients with irritable bowel syndrome: randomized placebo-controlled evidence. *Clin. Gastroenterol. Hepatol.* **6**, 765–771 (2008).
65. Treem, W. R., Ahsan, N., Kastoff, G. & Hyams, J. S. Fecal short-chain fatty acids in patients with diarrhea-predominant irritable bowel syndrome: *in vitro* studies of carbohydrate fermentation. *J. Pediatr. Gastroenterol. Nutr.* **23**, 280–286 (1996).
66. Mortensen, P. B., Andersen, J. R., Arffmann, S. & Krag, E. Short-chain fatty acids and the irritable bowel syndrome: the effect of wheat bran. *Scand. J. Gastroenterol.* **22**, 185–192 (1987).
67. Arpaia, N. *et al.* Metabolites produced by commensal bacteria promote peripheral regulatory T-cell generation. *Nature* **504**, 451–455 (2013).
68. Spiller, R. C. Inflammation as a basis for functional GI disorders. *Best Pract. Res. Clin. Gastroenterol.* **18**, 641–661 (2004).
69. Bercik, P., Verdu, E. F. & Collins, S. M. Is irritable bowel syndrome a low-grade inflammatory bowel disease? *Gastroenterol. Clin. North Am.* **34**, 235–245 (2005).
70. Rajilić-Stojanović, M. & de Vos, W. M. The first 1000 cultured species of the human gastrointestinal microbiota. *FEMS Microbiol. Rev.* **38**, 996–1047 (2014).
71. Taras, D., Simmering, R., Collins, M. D., Lawson, P. A. & Blaut, M. Reclassification of *Eubacterium formicigenerans* Holdeman and Moore 1974 as *Dorea formicigenerans* gen. nov., comb. nov., and description of *Dorea longicatena* sp. nov., isolated from human faeces. *Int. J. Syst. Evol. Microbiol.* **52**, 423–428 (2002).
72. King, T. S., Elia, M. & Hunter, J. O. Abnormal colonic fermentation in irritable bowel syndrome. *Lancet* **352**, 1187–1189 (1998).
73. Pritchard, S. E. *et al.* Fasting and postprandial volumes of the undisturbed colon: normal values and changes in diarrhea-predominant irritable bowel syndrome measured using serial MRI. *Neurogastroenterol. Motil.* **26**, 124–130 (2014).
74. Pimentel, M., Gunsalus, R. P., Rao, S. S. C. & Zhang, H. Methanogens in human health and disease. *Am. J. Gastroenterol. Suppl.* **1**, 28–33 (2012).
75. Attaluri, A., Jackson, M., Valestin, J. & Rao, S. S. C. Methanogenic flora is associated with altered colonic transit but not stool characteristics in constipation without IBS. *Am. J. Gastroenterol.* **105**, 1407–1411 (2010).
76. Chatterjee, S., Park, S., Low, K., Kong, Y. & Pimentel, M. The degree of breath methane production in IBS correlates with the severity of constipation. *Am. J. Gastroenterol.* **102**, 837–841 (2007).
77. Tooth, D. *et al.* Characterisation of faecal protease activity in irritable bowel syndrome with diarrhoea: origin and effect of gut transit. *Gut* **63**, 753–760 (2014).
78. Ivanov, D. *et al.* A serpin from the gut bacterium *Bifidobacterium longum* inhibits eukaryotic elastase-like serine proteases. *J. Biol. Chem.* **281**, 17246–17252 (2006).
79. Kerckhoffs, A. P. M. *et al.* Lower *Bifidobacteria* counts in both duodenal mucosa-associated and fecal microbiota in irritable bowel syndrome patients. *World J. Gastroenterol.* **15**, 2887–2892 (2009).
80. Rajilić-Stojanović, M. Function of the microbiota. *Best Pract. Res. Clin. Gastroenterol.* **27**, 5–16 (2013).
81. Jørgensen, J. & Mortensen, P. Hydrogen sulfide and colonic epithelial metabolism. *Dig. Dis. Sci.* **46**, 1722–1732 (2001).
82. Thiennimitr, P. *et al.* Intestinal inflammation allows *Salmonella* to use ethanolamine to compete with the microbiota. *Proc. Natl Acad. Sci. USA* **108**, 17480–17485 (2011).
83. Weissfeld, A. S. & Sonnenwirth, A. C. Rapid isolation of *Yersinia* spp. from feces. *J. Clin. Microbiol.* **15**, 508–510 (1982).
84. Spiller, R. Serotonin and GI clinical disorders. *Neuropharmacology* **55**, 1072–1080 (2008).
85. Jeffery, I. B. *et al.* An irritable bowel syndrome subtype defined by species-specific alterations in faecal microbiota. *Gut* **61**, 997–1006 (2012).
86. Mayer, E. A. & Bushnell, M. C. in *Functional Pain Syndromes: Presentation and Pathophysiology* 531–565 (eds Mayer, E. A. & Bushnell, M. C.) (IASP Press, 2009).
87. Mayer, E. A. *et al.* Functional GI disorders: from animal models to drug development. *Gut* **57**, 384–404 (2008).
88. Camilleri, M. Peripheral mechanisms in irritable bowel syndrome. *N. Engl. J. Med.* **367**, 1626–1635 (2012).
89. Keita, A. V. & Soderholm, J. D. The intestinal barrier and its regulation by neuroimmune factors. *Neurogastroenterol. Motil.* **22**, 718–733 (2010).

© 2016 Macmillan Publishers Limited. All rights reserved.

90. Elenkov, I. J. & Chrousos, G. P. Stress system — organization, physiology and immunoregulation. *Neuroimmunomodulation* **13**, 257–267 (2006).
91. Mayer, E. A., Savidge, T. & Shulman, R. J. Brain–gut microbiome interactions and functional bowel disorders. *Gastroenterology* **146**, 1500–1512 (2014).
92. Mayer, E. A., Labus, J. S., Tillisch, K., Cole, S. W. & Baldi, P. Towards a systems view of IBS. *Nat. Rev. Gastroenterol. Hepatol.* **12**, 592–605 (2015).
This narrative review summarizes the current knowledge on brain networks in relation to genomic-, gastrointestinal-, immune- and gut microbiota-related parameters and develops a novel model to understand their importance in IBS.
93. Grupe, D. W. & Nitschke, J. B. Uncertainty and anticipation in anxiety: an integrated neurobiological and psychological perspective. *Nat. Rev. Neurosci.* **14**, 488–501 (2013).
94. Elsenbruch, S. *et al.* Patients with irritable bowel syndrome have altered emotional modulation of neural responses to visceral stimuli. *Gastroenterology* **139**, 1310–1319 (2010).
95. Lackner, J. M. & Gurtman, M. B. Pain catastrophizing and interpersonal problems: a circumplex analysis of the communal coping model. *Pain* **110**, 597–604 (2004).
96. Mayer, E. A., Gupta, A., Kilpatrick, L. A. & Hong, J. Y. Imaging brain mechanisms in chronic visceral pain. *Pain* **156**, S50–S63 (2015).
97. Labus, J. S. *et al.* Regional brain morphology is associated with gut microbial metabolites in irritable bowel syndrome (IBS). *Gastroenterology* **148**, S-142 (2015).
98. Mayer, E. A. *et al.* Correlation between gene expression profiles in peripheral blood mononuclear cells and structural and functional brain networks in chronic visceral pain. *Neuropsychopharmacology* **39**, S641–S642 (2014).
99. Orand, A. *et al.* Catecholaminergic gene polymorphisms are associated with GI symptoms and morphological brain changes in irritable bowel syndrome. *PLoS ONE* **10**, e0135910 (2015).
100. Erpelding, N., Moayedi, M. & Davis, K. D. Cortical thickness correlates of pain and temperature sensitivity. *Pain* **153**, 1602–1609 (2012).
101. Gazouli, M. *et al.* Lessons learned: resolving the enigma of genetic factors in irritable bowel syndrome. *Nat. Rev. Gastroenterol. Hepatol.* **13**, 77–87 (2015).
This narrative review summarizes the current knowledge about genetic and epigenetic findings in IBS and critically discusses the methodological limitations that have to be overcome to underline clinical relevance of the findings.
102. Zhou, Q., Souba, W. W., Croce, C. M. & Verne, G. N. MicroRNA-29a regulates intestinal membrane permeability in patients with irritable bowel syndrome. *Gut* **59**, 775–784 (2010).
103. Zhou, Q. *et al.* MicroRNA 29 targets nuclear factor-κB-repressing factor and claudin 1 to increase intestinal permeability. *Gastroenterology* **148**, 158–169.e8 (2015).
104. Holliday, E. G. *et al.* Genome-wide association study identifies two novel genomic regions in irritable bowel syndrome. *Am. J. Gastroenterol.* **109**, 770–772 (2014).
105. Fukudo, S. *et al.* Impact of serotonin transporter gene polymorphism on brain activation by colorectal distention. *NeuroImage* **47**, 946–951 (2009).
106. Kilpatrick, L. A. *et al.* The *HTR3A* polymorphism c.–42C > T is associated with amygdala responsiveness in patients with irritable bowel syndrome. *Gastroenterology* **140**, 1943–1951 (2011).
107. Wong, B. S. *et al.* A Klothoβ variant mediates protein stability and associates with colon transit in irritable bowel syndrome with diarrhea. *Gastroenterology* **140**, 1934–1942 (2011).
108. Camilleri, M. *et al.* Genetic variation in *GPBAR1* predisposes to quantitative changes in colonic transit and bile acid excretion. *Am. J. Physiol. Gastrointest. Liver Physiol.* **307**, G508–G516 (2014).
109. Rao, A. S. *et al.* Chenodeoxycholate in females with irritable bowel syndrome-constipation: a pharmacodynamic and pharmacogenetic analysis. *Gastroenterology* **139**, 1549–1558.e1 (2010).
110. Wong, B. S. *et al.* Pharmacogenetics of the effects of colesevelam on colonic transit in irritable bowel syndrome with diarrhea. *Dig. Dis. Sci.* **57**, 1222–1226 (2012).
111. Camilleri, M. *et al.* Effect of increased bile acid synthesis or fecal excretion in irritable bowel syndrome-diarrhea. *Am. J. Gastroenterol.* **109**, 1621–1630 (2014).
112. Ek, W. E. *et al.* Exploring the genetics of irritable bowel syndrome: a GWA study in the general population and replication in multinational case–control cohorts. *Gut* **64**, 1774–1782 (2014).
113. Martinez, C. *et al.* Differential expression of miRNAs in the jejunal mucosa of IBS-D is involved in intestinal epithelial barrier dysfunction through modulation of specific tight junction proteins. *United European Gastroenterol. J.* **3** (Suppl. 5), A88 (2015).
114. Zhou, Q. *et al.* Decreased miR-199 augments visceral pain in patients with IBS through translational upregulation of TRPV1. *Gut* http://dx.doi.org/10.1136/gutjnl-2013-306464 (2015).
115. Wohlfarth, C. *et al.* Impaired miRNA regulation may lead to altered 5-HT4 receptor levels in IBS-D patients. *United European Gastroenterol. J.* **1** (Suppl. 1), A27 (2013).
116. Kapeller, J. *et al.* First evidence for an association of a functional variant in the microRNA-510 target site of the serotonin receptor-type 3E gene with diarrhea predominant irritable bowel syndrome. *Hum. Mol. Genet.* **17**, 2967–2977 (2008).
117. Fourie, N. H. *et al.* Elevated circulating miR-150 and miR-342-3p in patients with irritable bowel syndrome. *Exp. Mol. Pathol.* **96**, 422–425 (2014).
118. Gheinani, A. H., Burkhard, F. C. & Monastyrskaya, K. Deciphering microRNA code in pain and inflammation: lessons from bladder pain syndrome. *Cell. Mol. Life Sci.* **70**, 3773–3789 (2013).
119. Longstreth, G. F. *et al.* Functional bowel disorders. *Gastroenterology* **130**, 1480–1491 (2006).
120. Spiegel, B. M., Farid, M., Esrailian, E., Talley, J. & Chang, L. Is irritable bowel syndrome a diagnosis of exclusion?: A survey of primary care providers, gastroenterologists, and IBS experts. *Am. J. Gastroenterol.* **105**, 848–858 (2010).
121. Sood, R., Law, G. R. & Ford, A. C. Diagnosis of IBS: symptoms, symptom-based criteria, biomarkers or 'psychomarkers'? *Nat. Rev. Gastroenterol. Hepatol.* **11**, 683–691 (2014).
This narrative review describes how — based on current knowledge — symptoms, symptom-based criteria, biomarkers and psychomarkers, and their combinations may enable the diagnosis of IBS.
122. Manning, A. P., Thompson, W. G., Heaton, K. W. & Morris, A. F. Towards positive diagnosis of the irritable bowel. *Br. Med. J.* **2**, 653–654 (1978).
123. Drossman, D. A., Thompson, W. G. & Talley, N. J. Identification of sub-groups of functional gastrointestinal disorders. *Gastroenterol. Int.* **3**, 159–172 (1990).
124. Thompson, W. G. *et al.* Functional bowel disorders and functional abdominal pain. *Gut* **45** (Suppl. 2), II43–II47 (1999).
125. Ford, A. C. *et al.* Validation of the Rome III criteria for the diagnosis of irritable bowel syndrome in secondary care. *Gastroenterology* **145**,1262–1270. e1 (2013).
126. Whitehead, W. E. & Drossman, D. A. Validation of symptom-based diagnostic criteria for irritable bowel syndrome: a critical review. *Am. J. Gastroenterol.* **105**, 814–820 (2010).
127. Pimentel, M. *et al.* New clinical method for distinguishing D-IBS from other gastrointestinal conditions causing diarrhea: the LA/IBS diagnostic strategy. *Dig. Dis. Sci.* **55**, 145–149 (2010).
128. Bohn, L., Storsrud, S., Tornblom, H., Bengtsson, U. & Simren, M. Self-reported food-related gastrointestinal symptoms in IBS are common and associated with more severe symptoms and reduced quality of life. *Am. J. Gastroenterol.* **108**, 634–641 (2013).
129. Ford, A. C., Marwaha, A., Lim, A. & Moayyedi, P. Systematic review and meta-analysis of the prevalence of irritable bowel syndrome in individuals with dyspepsia. *Clin. Gastroenterol. Hepatol.* **8**, 401–409 (2010).
130. Vandvik, P. O., Wilhelmsen, I., Ihlebaek, C. & Farup, P. G. Comorbidity of irritable bowel syndrome in general practice: a striking feature with clinical implications. *Aliment. Pharmacol. Ther.* **20**, 1195–1203 (2004).
131. Jerndal, P. *et al.* Gastrointestinal-specific anxiety: an important factor for severity of GI symptoms and quality of life in IBS. *Neurogastroenterol. Motil.* **22**, 646–e179 (2010).
132. Kruis, W. *et al.* A diagnostic score for the irritable bowel syndrome. Its value in the exclusion of organic disease. *Gastroenterology* **87**, 1–7 (1984).
133. Soh, J. S. *et al.* The diagnostic value of a digital rectal examination compared with high-resolution anorectal manometry in patients with chronic constipation and fecal incontinence. *Am. J. Gastroenterol.* **110**, 1197–1204 (2015).
134. Tantiphlachiva, K., Rao, P., Attaluri, A. & Rao, S. S. C. Digital rectal examination is a useful tool for identifying patients with dyssynergia. *Clin. Gastroenterol. Hepatol.* **8**, 955–960 (2010).
135. American College of Gastroenterology Task Force on Irritable Bowel Syndrome *et al.* An evidence-based position statement on the management of irritable bowel syndrome. *Am. J. Gastroenterol.* **104**, S1–S35 (2009).
136. Cash, B. D. *et al.* The prevalence of celiac disease among patients with nonconstipated irritable bowel syndrome is similar to controls. *Gastroenterology* **141**, 1187–1193 (2011).
137. Menees, S. B., Powell, C., Kurlander, J., Goel, A. & Chey, W. D. A meta-analysis of the utility of C-reactive protein, erythrocyte sedimentation rate, fecal calprotectin, and fecal lactoferrin to exclude inflammatory bowel disease in adults with IBS. *Am. J. Gastroenterol.* **110**, 444–454 (2015).
138. Dickinson, B. & Surawicz, C. M. Infectious diarrhea: an overview. *Curr. Gastroenterol. Rep.* **16**, 399 (2014).
139. Pimentel, M. *et al.* Development and validation of a biomarker for diarrhea-predominant irritable bowel syndrome in human subjects. *PLoS ONE* **10**, e0126438 (2015).
140. Jones, M. P. *et al.* A biomarker panel and psychological morbidity differentiates the irritable bowel syndrome from health and provides novel pathophysiological leads. *Aliment. Pharmacol. Ther.* **39**, 426–437 (2014).
141. Vanner, S. J. *et al.* Predictive value of the Rome criteria for diagnosing the irritable bowel syndrome. *Am. J. Gastroenterol.* **94**, 2912–2917 (1999).
142. Whitehead, W. E. *et al.* Utility of red flag symptom exclusions in the diagnosis of irritable bowel syndrome. *Aliment. Pharmacol. Ther.* **24**, 137–146 (2006).
143. Stotzer, P. O. *et al.* Are the definitions for chronic diarrhoea adequate? Evaluation of two different definitions in patients with chronic diarrhoea. *United European Gastroenterol. J.* **3**, 381–386 (2015).
144. Gunnarsson, J. & Simren, M. Efficient diagnosis of suspected functional bowel disorders. *Nat. Clin. Pract. Gastroenterol. Hepatol.* **5**, 498–507 (2008).
145. Begtrup, L. M. *et al.* A positive diagnostic strategy is noninferior to a strategy of exclusion for patients with irritable bowel syndrome. *Clin. Gastroenterol. Hepatol.* **11**, 956–962.e1 (2013).
146. Spiegel, B. M. *et al.* Is a negative colonoscopy associated with reassurance or improved health-related quality of life in irritable bowel syndrome? *Gastrointest. Endosc.* **62**, 892–899 (2005).
147. Cairns, S. R. *et al.* Guidelines for colorectal cancer screening and surveillance in moderate and high risk groups. *Gut* **59**, 666–689 (2010).
148. Rex, D. K. *et al.* American College of Gastroenterology guidelines for colorectal cancer screening 2009 [corrected]. *Am. J. Gastroenterol.* **104**, 739–750 (2009).
149. Chey, W. D. *et al.* The yield of colonoscopy in patients with non-constipated irritable bowel syndrome: results from a prospective, controlled US trial. *Am. J. Gastroenterol.* **105**, 859–865 (2010).
150. Vijayvargiya, P., Camilleri, M., Shin, A. & Saenger, A. Methods for diagnosis of bile acid malabsorption in clinical practice. *Clin. Gastroenterol. Hepatol.* **11**, 1232–1239 (2013).
151. Money, M. E. & Camilleri, M. Review: management of postprandial diarrhea syndrome. *Am. J. Med.* **125**, 538–544 (2012).
152. Melchior, C., Gourcerol, G., Dechelotte, P., Leroi, A. M. & Ducrotte, P. Symptomatic fructose malabsorption in irritable bowel syndrome: a prospective study. *United European Gastroenterol. J.* **2**, 131–137 (2014).
153. Misselwitz, B. *et al.* Lactose malabsorption and intolerance: pathogenesis, diagnosis and treatment. *United European Gastroenterol. J.* **1**, 151–159 (2013).
154. Rana, S. V. & Malik, A. Breath tests and irritable bowel syndrome. *World J. Gastroenterol.* **20**, 7587–7601 (2014).
155. Simren, M. & Stotzer, P. O. Use and abuse of hydrogen breath tests. *Gut* **55**, 297–303 (2006).
156. Ford, A. C., Spiegel, B. M., Talley, N. J. & Moayyedi, P. Small intestinal bacterial overgrowth in irritable bowel syndrome: systematic review and meta-analysis. *Clin. Gastroenterol. Hepatol.* **7**, 1279–1286 (2009).
157. Posserud, I., Stotzer, P. O., Bjornsson, E. S., Abrahamsson, H. & Simren, M. Small intestinal bacterial overgrowth in patients with irritable bowel syndrome. *Gut* **56**, 802–808 (2007).

© 2016 Macmillan Publishers Limited. All rights reserved.

158. Heaton, K. W. *et al.* Symptoms of irritable bowel syndrome in a British urban community: consulters and nonconsulters. *Gastroenterology* **102**, 1962–1967 (1992).
159. Ringstrom, G., Abrahamsson, H., Strid, H. & Simren, M. Why do subjects with irritable bowel syndrome seek health care for their symptoms? *Scand. J. Gastroenterol.* **42**, 1194–1203 (2007).
160. Flik, C. E., Laan, W., Smout, A. J., Weusten, B. L. & de Wit, N. J. Comparison of medical costs generated by IBS patients in primary and secondary care in the Netherlands. *BMC Gastroenterol.* **15**, 168 (2015).
161. Khan, S. & Chang, L. Diagnosis and management of IBS. *Nat. Rev. Gastroenterol. Hepatol.* **7**, 565–581 (2010).
162. Hausteiner-Wiehle, C. & Henningsen, P. Irritable bowel syndrome: relations with functional, mental, and somatoform disorders. *World J. Gastroenterol.* **20**, 6024–6030 (2014).
163. Posserud, I. *et al.* Symptom pattern following a meal challenge test in patients with irritable bowel syndrome and healthy controls. *United European Gastroenterol. J.* **1**, 358–367 (2013).
164. Bohn, L., Storsrud, S. & Simren, M. Nutrient intake in patients with irritable bowel syndrome compared with the general population. *Neurogastroenterol. Motil.* **25**, 23–30.e21 (2013).
165. Hayes, P. A., Fraher, M. H. & Quigley, E. M. Irritable bowel syndrome: the role of food in pathogenesis and management. *Gastroenterol. Hepatol (N. Y.)* **10**, 164–174 (2014).
166. Wahnschaffe, U., Ullrich, R., Riecken, E. O. & Schulzke, J. D. Celiac disease-like abnormalities in a subgroup of patients with irritable bowel syndrome. *Gastroenterology* **121**, 1329–1338 (2001).
167. Kellow, J. E., Miller, L. J., Phillips, S. F., Zinsmeister, A. R. & Charboneau, J. W. Altered sensitivity of the gallbladder to cholecystokinin octapeptide in irritable bowel syndrome. *Am. J. Physiol.* **253**, G650–G655 (1987).
168. Kennedy, P. J. *et al.* Acute tryptophan depletion reduces kynurenine levels: implications for treatment of impaired visuospatial memory performance in irritable bowel syndrome. *Psychopharmacology (Berl.)* **232**, 1357–1371 (2015).
169. Ford, A. C. *et al.* American College of Gastroenterology monograph on the management of irritable bowel syndrome and chronic idiopathic constipation. *Am. J. Gastroenterol.* **109**, S2–S26 (2014).
170. Guo, Y. B. *et al.* Association between diet and lifestyle habits and irritable bowel syndrome: a case–control study. *Gut Liver* **9**, 649–656 (2014).
171. Halmos, E. P., Power, V. A., Shepherd, S. J., Gibson, P. R. & Muir, J. G. A diet low in FODMAPs reduces symptoms of irritable bowel syndrome. *Gastroenterology* **146**, 67–75.e5 (2014).
172. Bohn, L. *et al.* Diet low in FODMAPs reduces symptoms of irritable bowel syndrome as well as traditional dietary advice: a randomized controlled trial. *Gastroenterology* **149**, 1399–1407.e2 (2015).
In this paper, the data compare a diet low in FODMAPs with conventional dietary advice in IBS and report equal efficacy, thus counteracts the current enthusiasm with low FODMAP diets in patients with IBS.
173. Moayyedi, P. *et al.* The effect of dietary intervention on irritable bowel syndrome: a systematic review. *Clin. Transl Gastroenterol.* **6**, e107 (2015).
174. Halmos, E. P. *et al.* Diets that differ in their FODMAP content alter the colonic luminal microenvironment. *Gut* **64**, 93–100 (2015).
175. Berg, L. K. *et al.* Effect of fructose-reduced diet in patients with irritable bowel syndrome, and its correlation to a standard fructose breath test. *Scand. J. Gastroenterol.* **48**, 936–943 (2013).
176. Wilder-Smith, C. H., Materna, A., Wermelinger, C. & Schuler, J. Fructose and lactose intolerance and malabsorption testing: the relationship with symptoms in functional gastrointestinal disorders. *Aliment. Pharmacol. Ther.* **37**, 1074–1083 (2013).
177. Ludvigsson, J. F. *et al.* The Oslo definitions for coeliac disease and related terms. *Gut* **62**, 43–52 (2013).
178. Vazquez-Roque, M. I. *et al.* A controlled trial of gluten-free diet in patients with irritable bowel syndrome-diarrhea: effects on bowel frequency and intestinal function. *Gastroenterology* **144**, 903–911.e3 (2013).
179. Rodrigo, L., Blanco, I., Bobes, J. & de Serres, F. J. Effect of one year of a gluten-free diet on the clinical evolution of irritable bowel syndrome plus fibromyalgia in patients with associated lymphocytic enteritis: a case–control study. *Arthritis Res. Ther.* **16**, 421 (2014).
180. Biesiekierski, J. R. *et al.* No effects of gluten in patients with self-reported non-celiac gluten sensitivity after dietary reduction of fermentable, poorly absorbed, short-chain carbohydrates. *Gastroenterology* **145**, 320–328.e3 (2013).
181. Lee, K. N. & Lee, O. Y. Intestinal microbiota in pathophysiology and management of irritable bowel syndrome. *World J. Gastroenterol.* **20**, 8886–8897 (2014).
182. Shanahan, F. & Quigley, E. M. Manipulation of the microbiota for treatment of IBS and IBD-challenges and controversies. *Gastroenterology* **146**, 1554–1563 (2014).
183. Ford, A. C. *et al.* Efficacy of prebiotics, probiotics, and synbiotics in irritable bowel syndrome and chronic idiopathic constipation: systematic review and meta-analysis. *Am. J. Gastroenterol.* **109**, 1547–1561 (2014).
184. Mazurak, N., Broelz, E., Storr, M. & Enck, P. Probiotic therapy in the irritable bowel syndrome — why is the evidence for clinical efficacy still poor and what can be done about it? *J. Neurogastroeterol. Motil.* **21**, 471–485 (2015).
185. Hill, C. *et al.* Expert consensus document. The International Scientific Association for Probiotics and Prebiotics consensus statement on the scope and appropriate use of the term probiotic. *Nat. Rev. Gastroenterol. Hepatol.* **11**, 506–514 (2014).
186. Krueger, D. *et al.* Effect of hyoscine butylbromide (Buscopan®) on cholinergic pathways in the human intestine. *Neurogastroenterol. Motil.* **25**, e530–e539 (2013).
187. Ford, A. C. *et al.* Effect of fibre, antispasmodics, and peppermint oil in the treatment of irritable bowel syndrome: systematic review and meta-analysis. *BMJ* **337**, a2313 (2008).
188. Ruepert, L. *et al.* Bulking agents, antispasmodics and antidepressants for the treatment of irritable bowel syndrome. *Cochrane Database Syst. Rev.* **8**, CD003460 (2011).
189. Cash, B. D., Epstein, M. S. & Shah, S. M. A novel delivery system of peppermint oil is an effective therapy for irritable bowel syndrome symptoms. *Dig. Dis. Sci.* **61**, 560–571 (2016).
190. National Institute of Health and Care Excellence. CG61. Irritable bowel syndrome in adults: diagnosis and management of irritable bowel syndrome in primary care. *NICE* [online], http://www.nice.org.uk/guidance/cg61/documents (2008).
191. Moret, C. & Briley, M. Antidepressants in the treatment of fibromyalgia. *Neuropsychiatr. Dis. Treat.* **2**, 537–548 (2006).
192. Castro, J. *et al.* Linaclotide inhibits colonic nociceptors and relieves abdominal pain via guanylate cyclase-C and extracellular cyclic guanosine 3′,5′-monophosphate. *Gastroenterology* **145**, 1334–1346.e11 (2013).
193. Chey, W. D. *et al.* Linaclotide for irritable bowel syndrome with constipation: a 26-week, randomized, double-blind, placebo-controlled trial to evaluate efficacy and safety. *Am. J. Gastroenterol.* **107**, 1702–1712 (2012).
194. Rao, S. *et al.* A 12-week, randomized, controlled trial with a 4-week randomized withdrawal period to evaluate the efficacy and safety of linaclotide in irritable bowel syndrome with constipation. *Am. J. Gastroenterol.* **107**, 1714–1724 (2012).
195. Drossman, D. A. *et al.* Clinical trial: lubiprostone in patients with constipation-associated irritable bowel syndrome — results of two randomized, placebo-controlled studies. *Aliment. Pharmacol. Ther.* **29**, 329–341 (2009).
196. Tack, J., van Outryve, M., Beyens, G., Kerstens, R. & Vandeplassche, L. Prucalopride (Resolor) in the treatment of severe chronic constipation in patients dissatisfied with laxatives. *Gut* **58**, 357–365 (2009).
197. Efskind, P. S., Bernklev, T. & Vatn, M. H. A double-blind placebo-controlled trial with loperamide in irritable bowel syndrome. *Scand. J. Gastroenterol.* **31**, 463–468 (1996).
198. Lembo, A. J. *et al.* Eluxadoline for irritable bowel syndrome with diarrhea. *N. Engl. J. Med.* **374**, 242–253 (2016).
199. Clave, P. Treatment of IBS-D with 5-HT3 receptor antagonists versus spasmolytic agents: similar therapeutical effects from heterogeneous pharmacological targets. *Neurogastroenterol. Motil.* **23**, 1051–1055 (2011).
200. Berman, S. M. *et al.* Condition-specific deactivation of brain regions by 5-HT3 receptor antagonist alosetron. *Gastroenterology* **123**, 969–977 (2002).
201. Chang, L. *et al.* Incidence of ischemic colitis and serious complications of constipation among patients using alosetron: systematic review of clinical trials and post-marketing surveillance data. *Am. J. Gastroenterol.* **101**, 1069–1079 (2006).
202. Garsed, K. *et al.* A randomised trial of ondansetron for the treatment of irritable bowel syndrome with diarrhoea. *Gut* **63**, 1617–1625 (2014).
203. Fukudo, S., Ida, M., Akiho, H., Nakashima, Y. & Matsueda, K. Effect of ramosetron on stool consistency in male patients with irritable bowel syndrome with diarrhea. *Clin. Gastroenterol. Hepatol.* **12**, 953–959.e4 (2014).
204. Pimentel, M. *et al.* Rifaximin therapy for patients with irritable bowel syndrome without constipation. *N. Engl. J. Med.* **364**, 22–32 (2011).
205. Didari, T., Mozaffari, S., Nikfar, S. & Abdollahi, M. Effectiveness of probiotics in irritable bowel syndrome: updated systematic review with meta-analysis. *World J. Gastroenterol.* **21**, 3072–3084 (2015).
206. Madisch, A., Holtmann, G., Plein, K. & Hotz, J. Treatment of irritable bowel syndrome with herbal preparations: results of a double-blind, randomized, placebo-controlled, multi-centre trial. *Aliment. Pharmacol. Ther.* **19**, 271–279 (2004).
207. Simmen, U. *et al.* Binding of STW 5 (Iberogast®) and its components to intestinal 5-HT, muscarinic M3, and opioid receptors. *Phytomedicine* **13** (Suppl. 5), 51–55 (2006).
208. Fond, G. *et al.* Anxiety and depression comorbidities in irritable bowel syndrome (IBS): a systematic review and meta-analysis. *Eur. Arch. Psychiatry Clin. Neurosci.* **264**, 651–660 (2014).
209. Hungin, A. P. *et al.* Irritable bowel syndrome: an integrated explanatory model for clinical practice. *Neurogastroenterol. Motil.* **27**, 750–763 (2015).
210. Drossman, D. A. *et al.* A focus group assessment of patient perspectives on irritable bowel syndrome and illness severity. *Dig. Dis. Sci.* **54**, 1532–1541 (2009).
211. Owens, D. M., Nelson, D. K. & Talley, N. J. The irritable bowel syndrome: long-term prognosis and the physician-patient interaction. *Ann. Intern. Med.* **122**, 107–112 (1995).
212. Spiller, R. *et al.* Guidelines on the irritable bowel syndrome: mechanisms and practical management. *Gut* **56**, 1770–1798 (2007).
213. Fukudo, S. *et al.* Evidence-based clinical practice guidelines for irritable bowel syndrome. *J. Gastroenterol.* **50**, 11–30 (2015).
214. Ford, A. C. *et al.* Effect of antidepressants and psychological therapies, including hypnotherapy, in irritable bowel syndrome: systematic review and meta-analysis. *Am. J. Gastroenterol.* **109**, 1350–1365 (2014).
215. Gaylord, S. A. *et al.* Mindfulness training reduces the severity of irritable bowel syndrome in women: results of a randomized controlled trial. *Am. J. Gastroenterol.* **106**, 1678–1688 (2011).
216. Ware, J. E. Jr & Sherbourne, C. D. The MOS 36-item Short-Form Health Survey (SF-36). I. Conceptual framework and item selection. *Med. Care* **30**, 473–483 (1992).
217. EuroQoL-Group. EuroQol — a new facility for the measurement of health-related quality of life. *Health Policy* **16**, 199–208 (1990).
218. Patrick, D. L., Drossman, D. A., Frederick, I. O., DiCesare, J. & Puder, K. L. Quality of life in persons with irritable bowel syndrome: development and validation of a new measure. *Dig. Dis. Sci.* **43**, 400–411 (1998).
219. Hahn, B. A., Kirchdoerfer, L. J., Fullerton, S. & Mayer, E. Evaluation of a new quality of life questionnaire for patients with irritable bowel syndrome. *Aliment. Pharmacol. Ther.* **11**, 547–552 (1997).
220. Gralnek, I. M., Hays, R. D., Kilbourne, A., Naliboff, B. & Mayer, E. A. The impact of irritable bowel syndrome on health-related quality of life. *Gastroenterology* **119**, 654–660 (2000).
221. Halder, S. L. *et al.* Impact of functional gastrointestinal disorders on health-related quality of life: a population-based case–control study. *Aliment. Pharmacol. Ther.* **19**, 233–242 (2004).
This population-based, nested case–control study compares patients with IBS, patients with functional dyspepsia and patients with both diagnoses to healthy control individuals for health-related QOL (SF-36) and reports that most of the determinants of QOL are psychological in nature.

© 2016 Macmillan Publishers Limited. All rights reserved.

222. Singh, P. *et al.* Patients with irritable bowel syndrome-diarrhea have lower disease-specific quality of life than irritable bowel syndrome-constipation. *World J. Gastroenterol.* **21**, 8103–8109 (2015).
223. Creed, F. *et al.* Health-related quality of life and health care costs in severe, refractory irritable bowel syndrome. *Ann. Intern. Med.* **134**, 860–868 (2001).
224. Spiegel, B. M. *et al.* Clinical determinants of health-related quality of life in patients with irritable bowel syndrome. *Arch. Intern. Med.* **164**, 1773–1780 (2004).
225. El-Serag, H. B., Olden, K. & Bjorkman, D. Health-related quality of life among persons with irritable bowel syndrome: a systematic review. *Aliment. Pharmacol. Ther.* **16**, 1171–1185 (2002).
226. Fukudo, S., Hongo, M., Kaneko, H., Takano, M. & Ueno, R. Lubiprostone increases spontaneous bowel movement frequency and quality of life in patients with chronic idiopathic constipation. *Clin. Gastroenterol. Hepatol.* **13**, 294–301.e5 (2015).
227. Fukudo, S., Hongo, M., Kaneko, H. & Ueno, R. Efficacy and safety of oral lubiprostone in constipated patients with or without irritable bowel syndrome: a randomized, placebo-controlled and dose-finding study. *Neurogastroenterol. Motil.* **23**, 544–e205 (2011).
228. Ford, A. C. *et al.* Efficacy of 5-HT3 antagonists and 5-HT4 agonists in irritable bowel syndrome: systematic review and meta-analysis. *Am. J. Gastroenterol.* **104**, 1831–1843 (2009).
229. Bardhan, K. D. *et al.* A double-blind, randomized, placebo-controlled dose-ranging study to evaluate the efficacy of alosetron in the treatment of irritable bowel syndrome. *Aliment. Pharmacol. Ther.* **14**, 23–34 (2000).
230. Camilleri, M. *et al.* Efficacy and safety of alosetron in women with irritable bowel syndrome: a randomised, placebo-controlled trial. *Lancet* **355**, 1035–1040 (2000).
231. Marciani, L. *et al.* Stimulation of colonic motility by oral PEG electrolyte bowel preparation assessed by MRI: comparison of split versus single dose. *Neurogastroenterol. Motil.* **26**, 1426–1436 (2014).
232. Malagelada, C. *et al.* Classification of functional bowel disorders by objective physiological criteria based on endoluminal image analysis. *Am. J. Physiol. Gastrointest. Liver Physiol.* **309**, G413–G419 (2015).
233. Rao, S. S. *et al.* Investigation of colonic and whole-gut transit with wireless motility capsule and radiopaque markers in constipation. *Clin. Gastroenterol. Hepatol.* **7**, 537–544 (2009).
234. Shiotani, A. *et al.* Pilot study of biomarkers for predicting effectiveness of ramosetron in diarrhea-predominant irritable bowel syndrome: expression of *S100A10* and polymorphisms of *TPH1*. *Neurogastroenterol. Motil.* **27**, 82–91 (2015).
235. Camilleri, M. *et al.* Serotonin-transporter polymorphism pharmacogenetics in diarrhea-predominant irritable bowel syndrome. *Gastroenterology* **123**, 425–432 (2002).
236. Barrett, J. C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.* **40**, 955–962 (2008).
237. Yang, J. *et al.* Lactose intolerance in irritable bowel syndrome patients with diarrhoea: the roles of anxiety, activation of the innate mucosal immune system and visceral sensitivity. *Aliment. Pharmacol. Ther.* **39**, 302–311 (2014).
238. Rodrigues, A. C., Nicholas Verne, G., Schmidt, S. & Mauderli, A. P. Hypersensitivity to cutaneous thermal nociceptive stimuli in irritable bowel syndrome. *Pain* **115**, 5–11 (2005).
239. Dorn, S. D. *et al.* Increased colonic pain sensitivity in irritable bowel syndrome is the result of an increased tendency to report pain rather than increased neurosensory sensitivity. *Gut* **56**, 1202–1209 (2007).
240. Christopher, N. L. & Bayless, T. M. Role of the small bowel and colon in lactose-induced diarrhea. *Gastroenterology* **60**, 845–852 (1971).
241. Marciani, L. *et al.* Postprandial changes in small bowel water content in healthy subjects and patients with irritable bowel syndrome. *Gastroenterology* **138**, 469–477.e1 (2010).
242. Florent, C. *et al.* Influence of chronic lactulose ingestion on the colonic metabolism of lactulose in man (an *in vivo* study). *J. Clin. Invest.* **75**, 608–613 (1985).
243. Vandeputte, D. *et al.* Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates. *Gut* **65**, 57–62 (2015).
244. Andresen, V. *et al.* Irritable bowel syndrome — the main recommendations. *Dtsch. Arztebl. Int.* **108**, 751–760 (2011).
245. Henrich, J. F. *et al.* Identifying effective techniques within psychological treatments for irritable bowel syndrome: a meta-analysis. *J. Psychosomat. Res.* **78**, 205–222 (2015).
This meta-analysis summarizes data from 48 RCTs, which tested psychological interventions in patients with IBS, and reports that, overall, improvement in gastrointestinal symptoms was strongly associated with improvement in psychological distress.
246. Creed, F. *et al.* The cost-effectiveness of psychotherapy and paroxetine for severe irritable bowel syndrome. *Gastroenterology* **124**, 303–317 (2003).
247. Peters, S. L., Muir, J. G. & Gibson, P. R. Review article: gut-directed hypnotherapy in the management of irritable bowel syndrome and inflammatory bowel disease. *Aliment. Pharmacol. Ther.* **41**, 1104–1115 (2015).
248. Hutton, J. Cognitive behaviour therapy for irritable bowel syndrome. *Eur. J. Gastroenterol. Hepatol.* **17**, 11–14 (2005).
249. Hyphantis, T., Guthrie, E., Tomenson, B. & Creed, F. Psychodynamic interpersonal therapy and improvement in interpersonal difficulties in people with severe irritable bowel syndrome. *Pain* **145**, 196–203 (2009).
250. Liegl, G., Plessen, C. Y., Leitner, A., Boeckle, M. & Pieh, C. Guided self-help interventions for irritable bowel syndrome: a systematic review and meta-analysis. *Eur. J. Gastroenterol. Hepatol.* **27**, 1209–1221 (2015).

Acknowledgements
The research leading to these results has received funding from the People Programme of the European Union's Seventh Framework Programme under REA grant agreement no. 607652 (NeuroGUT), involving some of the authors. Most of the authors are also engaged in and have received support of the international network COST Action BM1106 GENIEUR (GENes in Irritable Bowel Syndrome Research Network EURope, www.GENIEUR.eu).

Author contributions
Introduction (P.E. and R.C.S.); Epidemiology (P.E. and J.S.-K.); Mechanisms/pathophysiology (G.B., B.N., M.R.-S., M.Schemann and E.A.M.); Diagnosis, screening and prevention (M.Simren); Management (E.M.M.Q., Q.A., A.D.F. and S.Z.); Quality of life (S.F.); Outlook (R.C.S.); Overview of Primer (P.E.).

Competing interests
The authors have received unrestricted research grants from Alfa Wassermann (G.B.), Alimentary Health (E.M.M.Q.), Astellas (S.F.), Boehringer-Ingelheim (M.Schemann), Cadi Group (G.B.), Danone Research (M.Simren), Falk (G.B.), Ferring (M.Simren), IMA (G.B.), Italchimici (G.B.), Ironwood (R.C.S.), Kao (S.F.), Lesaffre (R.C.S.), Lotenzatto (G.B.), Ono Pharmaceuticals (Q.A. and S.F.), Parmalat (G.B.), Pfizer (Q.A.), Pretexin (Q.A.), Rhythme (E.M.M.Q.), Schwabe (M.Schemann), Shire (P.E.), Sofar (G.B.), Steigerwald (M.Schemann), SymbioPharm (P.E.), Therevance (E.M.M.Q.), Vivrant (E.M.M.Q.), Yakult (G.B.) and Zespri (G.B.), have served as a consultant/advisory board members for Actavis (E.M.M.Q.), Albireo (M.Simren), Afa Wassermann (G.B.), Alimentary Health (E.M.M.Q.), Allergan (G.B. and M.Simren), Almirall (A.D.F., E.M.M.Q., G.B., M.Simren, P.E. and R.C.S.), Astellas (S.F.), AstraZeneca (M.Simren and P.E.), Bionorika (M.Schemann), Boehringer-Ingelheim (P.E.), Chr Hansen (M.Simren), Commonwealth (E.M.M.Q. and G.B.), Danone (E.A.M., G.B., M.Simren and R.C.S.), Ferring (P.E.), Glycom (M.Simren), IM Sciences (E.M.M.Q.), Ironwood (G.B.), Italchimici (G.B.), Ipsen (R.C.S.), Kissei (S.F.), Melasci (G.B.), Menarini (G.B.), Noos (G.B.), Parmalat (G.B.), Nestlé (M.Simren), Salix (E.M.M.Q.), Shire (E.M.M.Q., M.Simren and P.E.), Sofar (G.B.), SymbioPharm (P.E.), Synergy (E.M.M.Q. and G.B.), Takeda (M.Schemann), Yakult (G.B.), Yuhan Corp (R.C.S.) and Zenspri (G.B.), and as a speaker for Abbott (S.F.), Alfa Wassermann (G.B.), Almirall (B.N., E.M.M.Q., G.B., M.Simren, P.E. and Q.A.), Astellas (S.F.), AstraZeneca (E.M.M.Q.), Chiesi (P.E.), Danone (G.B.), Falk (B.N. and P.E.), Gruenenthal (P.E. and Q.A.), Heel (P.E.), Ironwood (E.M.M.Q.), Melasci (G.B.), Menarini (G.B. and R.C.S.), Noos (G.B.), Shire (A.D.F., E.M.M.Q., G.B., M.Simren and P.E.), Sofar (G.B.), Steigerwald (M.Schemann), SymbioPharm (P.E.), Takeda (M.Simren), Tillotts (M.Simren) and Yakult (G.B.). J.S.-K., M.R.-S. and S.Z. declare no potential conflict of interests.

SUPPLEMENTARY INFORMATION
See online article: S1 (table)
ALL LINKS ARE ACTIVE IN THE ONLINE PDF

© 2016 Macmillan Publishers Limited. All rights reserved.